Identification and Characterization of a New Class of Highly Conserved Non-Coding Sequences in Plants

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Konstantinos Kritsas

aus

Griechenland

Promotionskomitee

Prof. Dr. Ueli Grossniklaus (Vorsitz und Leitung der Dissertation)

PD Dr. Thomas Wicker

Prof. Dr. C-ting Wu

Zürich, 2014

To my grandmother Maria Brouzioti

Table of Contents

SUMMARY

Ultraconserved elements (UCEs), DNA sequences which are 100% identical between animal genomes are enigmatic features whose function is not well understood. UCEs are under strong purifying selection and a number of biological functions have been proposed to explain their robust conservation such as gene regulation, RNA processing and maintain genome integrity. However, all these functions are evolutionary tolerant to DNA sequence divergence without affecting their sequence specific interactions.

Here, we report the identification and characterization of highly conserved noncoding sequences in plant genomes. We have identified them after whole genome comparison studies between *Arabidopsis thaliana* (mouse-ear cress) and Vitis vinifera (grapevine). *Arabidopsis* and *Vitis* have diverged from their common ancestor ~115 Mya allowing significant changes at the DNA sequence to occur. We found 36 ULEs, which are >55 bp long and share at least 85% sequence identity. Interestingly, these elements exhibit properties similar to the mammalian UCEs, such that we named them UCE-like elements (ULEs). In addition to sequence constraints our data indicate that ULEs are functional elements. Further analysis showed that ULEs are under strong purifying selection. All of them have a sharp drop of A-T content just at their borders, and they are enriched next to genes involved in development. Intriguingly, the latter show preferential expression in undifferentiated cells. By comparing the genomes of *Brachypodium distachyon* (purple false brome) and *Oryza sativa* (rice), species that diverged ~50 Mya, we identified a different set of ULEs with similar properties in monocotyledons.

Surprisingly, likewise their animal counterparts, ULEs are depleted from segmental duplications. This observation led to the suggestion that ULEs or the regions that contain them are dosage sensitive. Our hypothesis about the function of ULEs is that they serve as agents of chromosome copy counting. According to this, the two homologous ULEs may compare their sequence perhaps through chromosome pairing to ensure the exact number of chromosomes. We employed a cytogenetic approach, fluorescence *in situ* hybridization (FISH) and found evidence that ULE regions exhibit increased pairing frequency in somatic cells relative to regions that are depleted from ULEs. We further

investigated the potential dosage-sensitivity nature of ULEs. Perturbation of one ULE resulted in distorted transmission efficiency of the corresponding allele in the offspring. Conversely, transmission efficiency of the same mutant was not distorted in an aneuploid context. Moreover, addition of four extra copies of ULEs did not yield any obvious phenotypes. Further investigation remains necessary to confirm a general role of ULEs in surveying genome dosage and integrity.

ZUSAMMENFASSUNG

Ultraconserved elements (UCEs) sind DNA-Sequenzen, die 100% Sequenzidentität zwischen verschiedenen tierischen Genomen aufweisen. Die Funktion dieser Elemente ist bisher ungeklärt. UCEs unterliegen starker, negativer Selektion, und es wurden verschiedene biologische Funktionen wie Genregulation, RNA-Prozessierung und Erhaltung der genomischen Integrität vorgeschlagen. Allerdings sind all diese Funktionen evolutionär tolerant gegenüber Sequenzdivergenz, ohne dass ihre sequenzspezifischen Interaktionen dadurch beeinflusst würden.

In dieser Arbeit beschreiben wir die Identifizierung und Charakterisierung hochkonservierter, nicht-kodierender Sequenzen in pflanzlichen Genomen. Diese wurden durch den Vergleich der kompletten Genome von *Arabidopsis thaliana* (Ackerschmalwand) und *Vitis vinifera* (Weinrebe) identifiziert. Die phylogenetische Trennung von *Arabidopsis* und *Vitis* von ihrem gemeinsamen Vorfahren fand vor ca. 115 Mio. Jahren statt, eine Zeitspanne, die signifikante Veränderungen der DNA-Sequenz zugelassen hat. Wir haben 36 UCE-ähnliche Elemente gefunden, welche >55 bp lang sind und mindestens 85% Sequenzidentität aufweisen. Interessanterweise zeigen diese Elemente ähnliche Eigenschaften wie die UCEs von Säugetieren, weshalb wir sie als UCE-like elements (ULEs) bezeichnet haben. Zusätzlich zu den Eigenschaften auf Sequenzebene zeigen unsere Daten, dass die ULEs funktionell sind. Weitere Analysen weisen darauf hin, dass die ULEs starker, negativer Selektion unterliegen. Alle ULEs weisen an ihren Enden einen starken Abfall im A-T-Gehalt auf und sie treten gehäuft neben Genen auf, die in der Entwicklung relevant sind. Bemerkenswerterweise werden diese Gene bevorzugt in undifferenzierten Zellen exprimiert. Durch den Genomvergleich zwischen *Brachypodium distachyon* (Zweiährige Zwenke) und *Oryza sativa* (Reis), deren phylogenetische Trennung vor ca. 50 Mio. Jahren stattfand, haben wir ein zusätzliches Set von ULEs mit ähnlichen Eigenschaften in Monokotyledonen gefunden.

Überraschenderweise waren ULEs, wie ihre tierischen Pendants, in "segmental duplications" nicht präsent. Diese Beobachtung suggeriert, dass die Anzahl Kopien der ULEs — oder der Regionen, die diese beinhalten — eine wichtige Rolle spielt. Wir

stellen die Hypothese auf, dass ULEs ein Mittel zur Zählung der Chromosomenzahl sind. Möglicherweise vergleichen die ULEs via Chromosomenpaarung ihre Sequenz, um die exakte Anzahl an Chromosomen sicherzustellen. Durch die Verwendung eines zytogenetischen Ansatzes, Fluoreszenz-*in-situ*-Hybridisierung (FISH), konnten wir nachweisen, dass Regionen mit ULEs in somatischen Zellen eine höhere Paarungsfrequenz aufweisen als Regionen ohne ULEs. Desweiteren haben wir die potentielle Kopienzahlsensitivität der ULEs untersucht. Die Mutation eines ULE resultierte in einer gestörten Transmissionseffizienz des entsprechenden Allels in der folgenden Generation. Umgekehrt hatte dieselbe Mutation im aneuploiden Kontext keinen Effekt auf die Transmissionseffizienz. Ebenso zeigte die Einführung von vier zusätzlichen ULE-Kopien keinen offensichtlichen Phänotyp. Weitere Untersuchungen sind nötig, um eine generelle Rolle der ULEs in der Überwachung der Chromosomenkopienzahl und der genomischen Integrität zu bestätigen.

CHAPTER 1

General Introduction

A brief history of genomics

The word genome comes from the combination of the German word *gen,* for gene, and the Greek suffix *–om,* from soma, *genom.* It was first introduced in 1920 from Winkler to describe the haploid set of chromosomes together with their genes, which define the foundation of each organism (Winkler 1920). The term genomics was proposed much later in 1986 from the geneticist Thomas H. Roderick. He came up with this term in an attempt to name the forthcoming at that time journal *Genomics* (McKusick 1997). Genomics is used to describe the sequencing methods and bioinformatic tools applied to determine and analyze the DNA sequence of an entire organism.

The first entire genome to be sequenced dates back in 1977, when Frederic Sanger and his team sequenced the genome of the bacteriophage φX174 (5’375 bp). In 1995, the first sequenced genome of a living organism was the bacterium *Haemophilus influenza* (1.8 Mb) (Fleischmann et al. 1995). A year after, the complete set of DNA of the first eukaryote *Saccharomyces cerevisiae* (12.1 Mb) was determined (Goffeau et al. 1996). But one of the biggest breakthroughs in genomics was the near complete sequencing of the 3.2 Gb of human genome (Lander et al. 2001; Venter et al. 2001). Since then, new sequencing technologies and computational tools have emerged allowing the cost of sequencing an entire genome to decrease substantially. Today, there is an ongoing explosion of organisms whose genomes are sequenced and assembled. There are more than 3’500 bacterial genomes and approximately 200 eukaryotic ones (protists, fungi, plants, insects, vertebrates), publicly available now.

Thus, genomics has become a whole new field of biology whose implications has been already fruitful in areas such as human diseases and drug discovery (Stankiewicz and Lupski 2010; Kramer and Cohen 2004). In agriculture, genomics are used to create disease-, pest- and drought-resistant crops, improve the health of farm animals and understand biodiversity (Zheng et al. 2003; Huang et al. 2005; Womack 2005; Schranz et al. 2007; Moyle 2008). Since genomes of very different species are available, there is the opportunity to compare their DNA sequences and learn more about the evolutionary history of modern species, as well as learn more about genetic elements which are essential for their living.

Comparative genomics

Fully sequenced genome comparisons between different species is an effective tool to address broader questions of evolutionary biology such as the mechanisms underlying genome evolution, understand the evolutionary forces that shape speciation and the phenotypic differences between closely related taxa. It also provides opportunities to identify genomic regions that distinguish one species from another. Comparative genomics can also be used to tackle practical problems such as annotation of previously undefined genes and infer phylogenies.

In addition, one of the major challenges in whole genome comparison studies is to determine the parts of the genome that are functional. The principle behind this is that over evolutionary time random mutations are eliminated from functional sequences due to negative selection, whereas non-functional sequences diverge in such degree that is almost impossible to identify them in other species (neutral selection). Thus, sequences which are evolving significant slower between different species are likely to have crucial functional roles.

But which genomes are appropriate for genome comparison? Sorting functional from non-functional DNA with comparative studies depends a lot on the genomes that are compared. It has been recommended to compare genomes that are derived from a common ancestor and which have diverged in such degree that significant amount of mutations has accumulated and selection has occurred (Ureta-Vidal et al. 2003). However, if genomes have diverged long enough, subsequently it is more challenging to identify common orthologous conserved regions. A balance should always be kept, too much similarity between genomes obscures the identification of functional elements from neutrally evolving sequences and too much divergence makes them invisible.

For example, in vertebrates the rate of divergence is 0.1-0.5% per million years. Thus, the 80 million years of evolutionary time since humans and mice diverged from their common ancestor is sufficient to identify functional orthologous sequences (Tautz 2000). In *Drosophila* species the evolution rate is higher, 2% per million years, which make 40 million years of evolutionary distance sufficient time to define conserved sequences (Tautz 2000).

Furthermore, the best choice for genomes to compare depends more on the biological questions are due to address. Tracking down allelic variants, such as single nucleotide polymorphisms (SNPs), small deletions or insertions, copy number variations (CNVs), which are responsible for phenotypic variation within a population, it is recommended

that genomes of the same species should be analyzed (Redon et al. 2006; Cao et al. 2011). To identify the genomic differences that makes humans different from their closest evolutionary relatives, the genomes of humans and chimpanzees need to be compared, even though these species diverged just 5-7 million years ago (Varki and Altheide 2005).

Comparative genomics enable researchers to shift some of their experiments from working with animal models such as mice and fish to more simple and less controversial organisms such as flies or even algae like *Chlamydomonas*. In a study conducted few years ago researchers kept the common proteins shared between *Chlamydomonas* and human after excluding those found in the non-flagellated proteome of *Arabidopsis*. This enabled them to perform functional assays on the simpler organism *Chlamydomonas* and identify a new gene responsible for Bardet-Biedl syndrome, a human ciliation disorder (Li et al. 2004).

Limitations of comparative genomics

Despite the vast resources of sequenced genomes, they are only valuable unless they are thoroughly and accurately annotated. Genome annotation is the process of attaching biological information to a sequence (Stein 2001). Although there is a plethora of annotation software, still it is computationally challenging to annotate low quality sequences like repetitive elements, map segmental duplications and identify variations like SNPs. In addition, many gene models suffer from errors in coding sequence definition. To address this issue, *in silico* annotation of sequences is getting corroborated with data from genome-wide transcriptomic, proteomic and high-throughput sequencing assays (Saha et al. 2002; Ossowski et al. 2008; Castellana et al. 2008; Schauer et al. unpublished). Curators of genomic databases are aware of these issues, therefore often revise genome annotations and release new versions of the genome.

One of the main findings of human – mouse comparisons was that 40% of the human genome can be aligned to that of mouse. On the other hand approximately 5% of the human genome is under purifying selection. This observation indicates that there is a portion of conserved sequence which likely is not functional (Waterston et al. 2002). To make it less probable that common sequences between genomes are not conserved by chance, they should also be present in other related species. Thus, increasing the number of species is one way to assure the identification of functional sequences (Boffelli et al. 2004). Another way to track functional sequences is to use species that are evolutionary

more distantly related than human and mouse. Genome comparisons between human and the teleost pufferfish *Fugus rubripes*, which are 450 million years apart, resulted the discovery of 1'000 putative human genes, which have not been described before (Aparicio et al. 2002).

Plant genomes

The number of sequenced genomes within the green lineage is continuously increasing. Until now the sequences of more than 90 plant genomes are available. A catalog of information resources for sequenced plant genomes is provided on table 1. In 2000, the draft sequence for the model system in plant research *Arabidopsis thaliana* was released (The *Arabidopsis* Genome Initiative 2000). Since then, the list of sequenced genomes expanded to important plants for research and agriculture. The genomes of *Medicago truncatula* and *Lotus japonicus* are important for elucidating the evolution of rhizobial symbioses; poplar (Populus trichocarpa) is used as a model species for forest research; *Brachypodium distachyon* is the model plant for grass research. More is to be discovered on the evolution of landplants, by knowing the genome sequence of moss (*Physcomitrella patens*), one of the first plants that conquered dry land. The genome of the green alga (*Chlamydomonas reinhardtii*) is used as a model system for studying photosynthesis and eukaryotic flagella (cilia) development.

One of the long standing promises of plant genomics is to transfer knowledge gained from model systems to agronomically important plants. In line with this effort, the genomes of the most important crops for food production; maize (*Zea mays*), rice (*Oryza sativa*), sorghum (*Sorghum bicolor*), barley (*Hordeum vulgare*), tomato (*Solanum lycopersicum*) and soybean (*Glycine max*) have been sequenced as well as grape (*Vitis vinifera*), cacao (*Theobroma cacao*) and cotton (*Gossypium raimondii*). However, plant genomes have distinct characteristics which make them in some way special than other systems.

Dynamic nature of plant genomes

Plant comparative genomics pose unique challenges due to their unique genome structure. Plant genomes are more volatile than mammals (Coghlan et al. 2005). Genome size in angiosperms (flowering plants) varies many orders of magnitude even between closely related species. For instance, the genome size of *Brachypodium* is 272 Mb,

whereas the estimated size of wheat is colossal, estimated 17’000 Mb. The two species belong to the same family and diverged ~50 Mya from their common ancestor (The International *Brachypodium* Initiative 2010). Two of the driving forces that create these large genome discrepancies are whole genome duplication events and transposable elements.

Polyploidization and subsequent DNA loss (diploidization) had a greater role in angiosperm evolution that in other eukaryotes. Half of the plants studied so far have undergone recent whole genome duplication (WGD) events and all angiosperms have had one or more ancient WGDs (Bowers et al. 2003). In contrast, in vertebrates the last WGD event occurred ~500 Mya (Dehal and Boore 2005; Hufton et al. 2008). Thus, the frequency of polyploidy occurs much more often in plants than in vertebrates.

The *Arabidopsis* lineage had undergone two major WGDs, named as α and β events, which are estimated to occur ~24-40 Mya (Blanc et al. 2003; Jaillon et al. 2007). In addition, another WGD (γ event), arose before the divergence of monocot and dicot plants (Bowers et al. 2003). The grapevine genome has not undergone any recent duplication since the divergence with *Arabidopsis*. However, regions in its genome are found in triplicates suggesting that its genome sequence is a product of ancestral hexaploidization event (γ event) (Jaillon et al. 2007). A recent WGD event has been detected with *Populus* genome which coincides with its divergence from the *Arabidopsis* lineage (Tuskan et al. 2006). Grasses share a WGD event that occurred ~60-70 Mya, since then another duplication took place in maize ~10 Mya (Van de Peer et al. 2009).

Except for genome duplication events, major contributors to the plant genome expansion are the transposable elements (TEs). TEs affect the genome by their ability to move and replicate, consequently shaping the genome (Wicker et al. 2007). In most plant species the majority of DNA consists of TEs. As mentioned before, the wheat genome is 62-fold larger than of *Brachypodium*. This difference is mostly caused by the presence of TEs. In *Brachypodium*, TEs comprise 21.4% of the genome, whereas in wheat, they contribute more than 80% (The International *Brachypodium* Initiative 2010). Thus, transposons boost genome expansion, which in plants seem to occur in short evolutionary times. For instance, in angiosperms more than 80% of LTR-retrotransposon movements have occurred within the last 5 Mya (Bennetzen 2005).

Plants have devised mechanisms to counterbalance genome expansion because of TE amplification by eliminating them. One way to do this is by creating small deletions in TEs through illegitimate recombination. Over time, TEs with deletions become inactive

and eventually will not be able to replicate (Devos et al. 2002). A genome comparison study between *A. thaliana* and *A. lyrata* provides evidence that the difference in their genome size (*A. thaliana* is smaller) is due to hundreds of thousands of small deletions on TEs (Hu et al. 2011).

So far, it is unknown why plants differ so strongly in TE content and genome size. One explanation could be that the mechanisms of TE amplification and TE elimination are working in different ways in plants than in other systems. For example, although TEs in plants have a life span of few million years, in mammals L1 repeats have been retained for more than 75 Mya (Waterston et al. 2000). It is also possible that these mechanisms could be influenced by the external environment, as been shown that TEs are activated over heat stress (Pecinka et al. 2010; Ito et al. 2011).

The large amounts of TEs in plant genomes makes sequencing of these genomes challenging as it is difficult to assemble the repetitive reads, since they map in multiple positions. For that reason the first crop genomes to be sequenced were small and only very recently larger genomes have been released.

Synteny, collinearity and conservation in plant genomes

Synteny and conservation are important information in order to elucidate the evolutionary history of genomes and identify the functional elements that shape genomes in coding and noncoding regions. Syntenic are homologous DNA regions between genomes that tend to have the same genetic components. If additionally the order and orientation is the same then the regions are called collinear. So collinearity is a more specific form of synteny. In that sense, a conserved sequence between two or more genomes should lie in orthologous or collinear regions.

As divergence time between genomes increases synteny/collinearity is eroding by reasons that were explained earlier, genome duplications, TE amplifications and deletions. In plants synteny is less retained than in animals. For instance, approximately 35% of genes between maize and sorghum are non-collinear, even though they diverged just 12 Mya (Bennetzen 2005). In contrast, despite human and mouse being 80 Mya apart, they still share vast genomic blocks of collinear sequence (Mural et al. 2002). Another example which illustrates the high degree of variation in plants as opposed to mammals is that a pair of maize individuals differs at 10-fold more sites than two human individuals do.

Synteny retention depends on the divergence time between the genomes. In dicots, approximately 90% of the genome has remained syntenic and genes are highly conserved between *A. thaliana* and *A. lyrata*. Divergence time of the two species is ~10 Mya. However, only 50% of the *A. lyrata* genome can be aligned on to *A. thaliana* revealing the high amount of rearrangements and deletions already appear in such short evolutionary window (Hu et al. 2011). Synteny is also preserved by comparing *Arabidopsis* with species of the same family, the *Brassicaceae* (Parkin et al. 2005; Slotte et al. 2013). However, in longer evolutionary distances, synteny erodes substantially between *Arabidopsis* and two other dicot species *Vitis vinifera* (~115 Mya apart) and *Carica papaya* (~72 Mya apart) (Freeling et al. 2008).

In grasses, most of the genes are in syntenic blocks because all of them derived from a common ancestor ~70 Mya that had an ancient duplication (Goff et al. 2002). Compared to grasses, synteny is poor between dicots (*Arabidopsis*) and monocots (rice). The two groups diverged more than 200 Mya. Regions where gene order is still preserved are relatively small and disrupted by non-collinear genes. In addition, the level of conservation is significant lower, with global mean identity of *Arabidopsis* proteins to rice to be 49.5% (Liu et al. 2001; Goff et al. 2002).

Arabidopsis - grapevine, *Brachypodium* – rice for comparative genomics

Arabidopsis is an excellent system for use in comparative genomics studies because its genome is the best annotated among plants. In addition, its genome is small (~127 Mb), there is a wealthy collection of genetic tools and currently there is vivid research on this system. For plant standards, the content of repetitive DNA is very small, just 10% (The *Arabidopsis* Genome Initiative 2000). The genome of grapevine that is available derives from Pinot Noir, a bred that was successively selfed to reach near-full homozygosity (93%) (Jaillon et al. 2007). *Arabidopsis* and grapevine are dicotyledonous plants, which diverged from their common ancestor ~115 Mya. Thus, sufficient time for purifying selection is exercised upon their genomes.

The *Arabidopsis* genome is highly rearranged. It has gone through two recent genome duplications, and in the last 10 Mya since its divergence from *A. lyrata*, it has undergone chromosome fusions, bringing the chromosome number from n=8 to n=5 (Tang et al. 2008). In addition, there is growing evidence that the genome of *Arabidopsis* is growing smaller with many deletions taking place (Hu et al. 2011). On the other hand, grapevine during its history had only one polyploidization event shared among all dicots, hence has

a relative preserved ancestral genome structure and provides an independent lineage to trace collinear regions which are under selection (Huang et al. 2009). On this account, conserved sequences found between these very different genomes indicate functional importance.

Brachypodium and rice are excellent models for genome comparisons. *Brachypodium*, a member of the Pooideae family, is a diploid, inbred, annual grass with a small life cycle and compact genome. Because of its amenability to transformation, the mutant and germplasm collection it is thought to be the model plant for grasses. Grasses are the major source for food production and currently there is growing research to use these crops for renewable research. Rice is one of the most important cereals. More than 500 million tons are produced annually. One third of the population depends on rice for more than 50% caloric intake (Goff et al. 2002). *Brachypodium* and rice are evolving independently ~40-53 million years. Despite the long period of evolution, they share extensive synteny and there was no genome duplication event after their divergence.

What are conserved non-coding sequences?

Conserved non-coding sequences (CNSs) are orthologous stretches of DNA whose sequence remains retained between different species. CNSs do not code for proteins, they are not part of repetitive DNA and they are not a product of horizontal DNA transfer from organelles.

Comparisons between genomes of different species revealed that there is an abundant portion of CNSs. The level of conservation is at least 70% and is higher than the average level of neutral sequence conservation. CNSs are present in yeast, worms, insects, and vertebrate species (Frazer et al. 2001; Bergman and Kreitman 2001; Mural et al. 2002; Dermitzakis et al. 2002, 2003, 2004; Kellis et al. 2003; Stein et al. 2003; Siepel et al. 2005). It is noted; that CNSs are expanding from yeast to vertebrates in accordance to the genome size and general biological complexity. For example, in vertebrates 58% of the conserved sequences are CNSs, whereas in insects, worm and yeast less than 7% of them are CNSs (Siepel et al. 2005). Surprisingly, mammalian CNSs are more conserved than protein coding sequences and noncoding RNAs (ncRNAs) (Dermitzakis et al. 2003). CNSs are also present in plants but until this study, the focus of research was on identification of short CNSs flanking a small number of orthologous genes (Kaplinsky et al. 2002; Guo and Moose 2003; Inada et al. 2003; Bossolini et al. 2007). Thus, in plants there were no whole genome comparison studies aiming at the identification of CNSs.

Ultraconserved Elements

An astonishing finding from cross-species genome comparison studies was the identification of, Ultraconserved Elements (UCEs) which are 100% identical between mammalian genomes. UCEs are sequences of size more than 200 bp, which are found in orthologous regions of human, mouse and rat genomes (Bejerano et al. 2004). UCEs are also present in evolutionary more distant species. The majority of them appeared during tetrapod evolution (Stephen et al. 2008), however many of them were also present in the jawed vertebrate ancestor, reflecting that they are more than 530 million years old (Wang et al. 2009). A separate class of UCEs has been also identified between the insect species *Drosophila melanogaster*, *Drosophila pseudoobscura* and *Anopheles gambiae*. However, in insects UCEs are less frequent and are smaller in size than the vertebrate ones (Glazov et al. 2005).

Except for a few exons, UCEs are overlapping intronic and intergenic regions; these are the non-coding UCEs (ncUCEs). Exonic UCEs are enriched next to genes involved in RNA processing whereas ncUCEs are clustered next to specific functional groups of genes such as transcription factor and development related genes (Bejerano et al. 2004). Intriguingly, ncUCEs have some other unique properties. There is a significant difference in the A-T content between the ncUCEs and their flanking regions. The transition of flanking sequence to ncUCEs is marked by a sharp drop of A-T content just at the border (Chiang et al. 2008). In addition, ncUCEs are only found in one copy in the genome (Bejerano et al. 2004). It turns out that evolutionary forces kept ncUCEs in just one copy by depleting them from segmental duplications and copy number variants, which suggests that they are dosage sensitive (Derti et al. 2006; Chiang et al. 2008)

One of the principles of genome comparative studies is that sequences which show small variation between species encompass important biological functions. Indeed, it was shown that ncUCEs are not just cold mutational spots but are under strong purifying selection. Selection forces applied on them are even stronger than the ones on coding sequences, implying that any mutations on them are potentially deleterious (Chen et al. 2007; Katzman et al. 2007).

Although ncUCEs seem to be under functional constraints removal of four them from the mouse genome did not affect the fitness of the mice, in terms of growth, longevity, pathology, metabolism and fertility (Ahituv et al. 2007). To justify, this surprising result

it was suggested that the effect of deletion ncUCEs could be evident over longer generation times.

However, there are a number of studies providing evidence that indeed ncUCE are functional elements. There are reports suggesting that ncUCEs are involved in epigenetic regulation of genes. NcUCEs harbor chromatin marks which in turn affect the expression of genes involved in embryonic stem cell development (Bernstein et al. 2006; Lee et al. 2006). *In vivo,* transgenic assays on mice embryos indicate that ncUCEs are acting as tissue specific enhancers indicating that they act as transcription regulators of key genes during mammalian development (Poulin et al. 2005; Pennacchio et al. 2006; Visel et al. 2008). Some ncUCEs are associated with human diseases. A SNP in ncUCE affect the function of genes that cause human autism disorder (Poitras et al. 2010) and another ncUCE is thought to induce apoptosis in colon cancer cells (Calin et al. 2007). From the analysis of protein interactions with more than 100 mammalian ncUCEs, it was revealed that more than 400 proteins are binding on them, especially transcription factors and chromatin remodelers. Therefore, it is proposed that ncUCEs are an amalgam of high density overlapping binding sites (Viturawong et al. 2013).

Chromosome copy counting hypothesis

Even though, ncUCEs are implied to act as enhancers or regulatory elements. This alone cannot fully explain the ultraconservation and ultraselection of these elements. In nature, there are no examples of DNA-DNA, DNA-RNA, DNA-protein interactions that require absolute conservation. Still these interactions can tolerate some sequence variation without affecting their interactions. Thus, the true purpose of their existence still remains enigmatic.

As mentioned before, ncUCEs are not just single copy in the haploid genome but they were never present in segmental duplications (SDs) occurred 40 million years ago as well as in a number of copy number variant (CNV) data sets (Derti et al. 2006; Chiang et al. 2008). Taking into account that ncUCEs predate the existence of SDs and CNVs it is claimed that ncUCEs are dosage sensitive.

Therefore, an alternative role for the function of ncUCEs has been proposed. Because of the genome-wide distribution of ncUCEs together with their uniqueness, it is implied that ncUCEs are participating in a copy counting mechanism (Derti et al. 2006). According to this model, in a diploid cell ncUCEs from the two homologous chromosomes are pairing

and compare each other at the nucleotide level. In the event, that there is sequence variation of significant magnitude this would stimulate deleterious events. This model suggests that deletions or duplications of ncUCEs would be eliminated through lethality, segregation distortion or lower fitness (Derti et al. 2006). That would be a reason why; ncUCEs are depleted from SDs and CNVs. In addition, the copy counting hypothesis does not exclude the enhancer nature of some ncUCEs as enhancers are also mediating DNA-DNA interactions.

Three - dimensional organization of the genomes

Despite our knowledge of genome sequences, our understanding of how genome is functioning is still limited. Genes and non-coding DNA can be active on specific cell types and inactive in others (Bernstein et al. 2012). To discern the forces that regulate the activity of functional elements in cell specific manner, it is important to understand how the genome is organized in the nucleus.

The three – dimensional folding of the chromosomes has been extensively investigated with fluorescence *is situ* hybridization (FISH) techniques. The first application of FISH was in 1980 where an RNA molecule was conjugated with a fluorochrome and used as a probe to hybridize to the DNA target (Bauman et al. 1980). FISH is a tool to microscopically visualize and detect chromosome organization and dynamics especially during interphase nuclei. FISH utilizes labeled nucleic acid probes which are complementary to the target DNA or RNA sequence. The relative nuclear position of genes, genomic segments or whole chromosomes can be analyzed by hybridization of the probe to fixed nuclei. The hybridization signals can be later visualized by fluorescence microscopy. Chromosome interactions can also be identified with chromosome conformation capture-based methods. Here, cells are cross-linked with formaldehyde to link chromosome regions that are in close proximity (Dekker et al. 2002).

In contrast to earlier view, chromosomes are well organized in the nucleus. In interphase nuclei, chromosomes do not mix but rather occupy their own territory, called chromosome territories (CTs) (Cremer and Cremer 2001). However, there is growing evidence that chromosomes intermingle with each other at the boundaries of CTs, making it possible that some of the interactions are functional (Branco and Pombo 2006). Thus, the physical position of a gene or of any other functional element in the nucleus can affect its active or inactive state. This is true in mouse erythroid tissues where it was shown that co-regulated genes and their regulatory elements in the nucleus are clustering

together to form transcriptional factories, in order to optimize their transcription (Schoenfelder et al. 2010). On the other hand, in *Drosophila,* silenced genes are also clustering in nuclear space at specialized structures called polycomb bodies (Bantignies et al. 2011) or in human nuclei silenced genes are associated with the nuclear lamina (Guelen et al. 2008). These findings challenge the notion that transcription of genes is simply one dimensional.

In plants with large genomes (>5'000 Mb/1C) such as wheat, rye, barley interphase chromosomes have a Rabl orientation, meaning that centromeres and telomeres are clustered in the opposite poles of the nucleus. This is not true for plants with smaller genome such as maize, sorghum and *Arabidopsis* where centromeres are located randomly in the periphery of the nucleus and telomeres are clustered in the nucleolus (Dong and Jiang 1998; Fransz et al. 2002). Heterochromatin is organized in condensed chromocenters. In plants, chromosomes do also occupy distinct CTs (Pecinka et al. 2004). However, it seems that interactions between loci in homologous chromosomes occur at random except for the nucleolus organizing regions (Pecinka et al. 2004). In agreement with animal studies, gene expression is regulated at the three-dimensional level as well. The expression of *FLC,* a gene responsible for flowering time regulation is decreased under cold induction under the control of polycomb proteins (Gendall et al. 2001). It turns out that after cold treatment *FLC* transcripts are physically clustering into foci in the nucleus (Rosa et al. 2013).

Our understanding of nucleus architecture and its implication to gene regulation is now becoming clear. FISH and chromosome conformation capture approaches are expected to provide a high resolution chromosome interaction maps.

Aims of this study

In this study, we investigate whether conserved non-coding sequences with similar level of conservation to UCEs exist in plants. To address this, the genomes of *Arabidopsis* and grapevine and the genomes of *Brachypodium* and rice were used in genome comparison studies. We further ask whether the plant non-coding sequences exhibit properties similar to UCEs and whether they are as well under purifying selection.

Regarding the function of the plant UCEs we explore the chromosome copy counting hypothesis arguing that enhancer activity alone is no sufficient to explain their highly conserved nature. Plant systems provide unique opportunities for testing this hypothesis. *Arabidopsis* plants produce thousand of seeds making them ideal model system to detect

even small deviations in fitness. In addition, unlike mammalian systems, *Arabidopsis* can tolerate aneuploidy. Trisomics, diploid plants with one extra chromosome, are viable in *Arabidopsis*. Thus, it is possible to investigate the fitness discrepancies by deleting a plant UCE in this more sensitized genetic background.

Taking advantage of the cytogenetic tools available in *Arabidopsis* as well as the vast resource of biological information, it makes it an excellent model system to study the occurrence of UCEs in plants and decipher their functional properties.

Table 1. List of resources for obtaining and analyzing plant genomic sequences

Phytozome	plant genome resource, comparative genome *analysis*	phytozome.net	Goodstein et al. 2012
PlantGDB	plant genome resource, comparative genome *analysis*	plantgdb.org	Dong et al. 2004
NCBI	plant genome resource	ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html	
JCVI	plant genome resource	jcvi.org/cms/research/groups/plant-genomics	
iPlant Collaborative	cyberinfrastructure platform	iplantcollaborative.org	Goff et al. 2011
VISTA	Plant genome alignments	genome.lbl.gov/vista/	Frazer et al. 2004
Plaza	comparative genome analysis	bioinformatics.psb.ugent.be/plaza	Van Bel et al. 2012
MIPS	comparative genome analysis	mips.helmholtz-muenchen.de/plant/genomes.jsp	Nussbaumer et al. 2013
EnsemblPlants	plant genome resource	plants.ensembl.org/index.html	EMBL-EBI
TAIR	information resource for the model plant *Arabidopsis thaliana*	arabidopsis.org	Swarbreck et al. 2008
1001 Genomes	information resource for sequence variation of *A. thaliana* ecotypes	1001genomes.org	Ossowski et al. 2008
1001 Epigenomes	information resource for epigenetic variation of *A. thaliana* ecotypes	neomorph.salk.edu/1001_epigenomes.html	Schmitz et al. 2013
Gramene	information resource for grass species	gramene.org	Ware et al. 2002
Sol genomics	information resource for solanaceae species	solgenomics.net	

References

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio L a, Rubin EM. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol* **5**: e234.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia J-M, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, et al. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**: 1301–10.

The *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**: 796–815.

Bantignies F, Roure V, Comet I, Leblanc B, Schuettengruber B, Bonnet J, Tixier V, Mas A, Cavalli G. 2011. Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. *Cell* **144**: 214–26.

Bauman JGJ, Wiegant J, Borst P, van Duijn P. 1980. new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA. *Exp Cell Res* **128**: 485–490.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* **304**: 1321–5.

Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van de Peer Y, Vandepoele K. 2012. Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol* **158**: 590–600.

Bennetzen JL. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr Opin Genet Dev* **15**: 621–7.

Bergman CM, Kreitman M. 2001. Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. *Genome Res* **11**: 1335–45.

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**: 315–26.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res* **13**: 137–44.

Boffeli D, Nobrega MA, Rubin EM. 2004. Comparative genomics at the vertebrate extremes. *Nat Rev Genet.* **5**: 456-65.

Bossolini E, Wicker T, Knobel P a, Keller B. 2007. Comparison of orthologous loci from small grass genomes Brachypodium and rice: implications for wheat genomics and grass genome annotation. *Plant J* **49**: 704–17.

Bowers JE, Chapman BA, Rong J. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**: 433–438.

Branco MR, Pombo A. 2006. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol* **4**: e138.

Calin G a, Liu C, Ferracin M, Hyslop T, Spizzo R, Sevignani C, Fabbri M, Cimmino A, Lee EJ, Wojcik SE, et al. 2007. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* **12**: 215–29.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, et al. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet* **43**: 956–63.

Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, Briggs SP. 2008. Discovery and revision of Arabidopsis genes by proteogenomics. *Proc Natl Acad Sci U S A* **105**: 21034–8.

Chen CTL, Wang JC, Cohen B a. 2007. The strength of selection on ultraconserved elements in the human genome. *Am J Hum Genet* **80**: 692–704.

Chiang CWK, Derti A, Schwartz D, Chou MF, Hirschhorn JN, Wu C-T. 2008. Ultraconserved elements: analyses of dosage sensitivity, motifs and boundaries. *Genetics* **180**: 2277–93.

Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L. 2005. Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends Genet* **21**: 673–82.

Cremer T, Cremer C. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**: 292–301.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3**: e314.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science* **295**: 1306–11.

Dermitzakis ET, Kirkness E, Schwarz S, Birney E, Reymond A, Antonarakis SE. 2004. Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res* **14**: 852–9.

Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV, et al. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420**: 578–82.

Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, Antonarakis SE. 2003. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* **302**: 1033–5.

Derti A, Roth FP, Church GM, Wu C. 2006. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet* **38**: 1216–20.

Devos KM, Brown JKM, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res* **12**: 1075–9.

Dong F, Jiang J. 1998. Non-Rabl patterns of centromere and telomere distribution in the interphase nuclei of plant cells. *Chromosome Res* **6**: 551–8.

Dong Q, Schlueter SD, Brendel V. 2004. PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res* **32**: D354–9.

Fleischmann RD, Adams MD, White O, Clayton R a, Kirkness EF, Kerlavage a R, Bult CJ, Tomb JF, Dougherty B a, Merrick JM. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**: 496–512.

Fransz P, De Jong JH, Lysak M, Castiglione MR, Schubert I. 2002. Interphase chromosomes in Arabidopsis are organized as well defined chromocenters from which euchromatin loops emanate. *Proc Natl Acad Sci U S A* **99**: 14584–9.

Frazer KA, Sheehan JB, Stokowski RP, Chen X, Hosseini R, Cheng JF, Fodor SP, Cox DR, Patil N. 2001. Evolutionarily conserved sequences on human chromosome 21. *Genome Res* **11**: 1651–9.

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* **32**: W273-9.

Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. 2008. Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. *Genome Res* **18**: 1924–37.

Gendall a R, Levy YY, Wilson a, Dean C. 2001. The VERNALIZATION 2 gene mediates the epigenetic regulation of vernalization in Arabidopsis. *Cell* **107**: 525–35.

Glazov E a, Pheasant M, McGraw E a, Bejerano G, Mattick JS. 2005. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res* **15**: 800–8.

Goff S a, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* **296**: 92–100.

Goff S a, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, Matasci N, Wang L, Hanlon M, Lenards A, et al. 2011. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Front Plant Sci* **2**: 34.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. 1996. Life with 6000 Genes. *Science* **274**: 562–567.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**: D1178–86.

Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W, et al. 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**: 948–51.

Guo H, Moose SP. 2003. Conserved Noncoding Sequences among Cultivated Cereal Genomes Identify Candidate Regulatory Sequence Elements and Patterns of Promoter Evolution. *Plant Cell* **15**: 1143–1158.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, Fahlgren N, Fawcett J a, Grimwood J, Gundlach H, et al. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet* **43**: 476–81.

Huang S, van der Vossen EA, Kuang H, Vleeshouwers VG, Zhang N, Borm TJ, van Eck HJ, Baker B, Jacobsen E, Visser RG. 2005. *Plant J.* **42**: 251-61.

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, et al. 2009. The genome of the cucumber, Cucumis sativus L. *Nat Genet* **41**: 1275–81.

Hufton AL, Groth D, Vingron M, Lehrach H, Poustka AJ, Panopoulou G. 2008. Early vertebrate whole genome duplications were predated by a period of intense genome rearrangement. *Genome Res* **18**: 1582–91.

Inada DC, Bashir A, Lee C, Thomas BC, Ko C, Goff S a, Freeling M. 2003. Conserved noncoding sequences in the grasses. *Genome Res* **13**: 2030–41.

The International *Brachypodium* Initiative. 2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* **463**: 763–8.

Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. 2011. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature* **472**: 115–9.

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–7.

Kaplinsky NJ, Braun DM, Penterman J, Goff S a, Freeling M. 2002. Utility and distribution of conserved noncoding sequences in the grasses. *Proc Natl Acad Sci U S A* **99**: 6147–51.

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D. 2007. Human genome ultraconserved elements are ultraselected. *Science* **317**: 915.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241–54.

Kramer R, Cohen D. 2004. Functional genomics to new drug targets. *Nat Rev Drug Discov.* **3**: 965-72.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Lee TI, Jenner RG, Boyer L a, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K, et al. 2006. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**: 301–13.

Li JB, Gerdes JM, Haycraft CJ, Fan Y, Teslovich TM, May-Simera H, Li H, Blacque OE, Li L, Leitch CC, et al. 2004. Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. *Cell* **117**: 541–52.

Liu H, Sachidanandam R, Stein L. 2001. Comparative genomics between rice and Arabidopsis shows scant collinearity in gene order. *Genome Res* **11**: 2020–6.

McKusick V a. 1997. Genomics: structural and functional studies of genomes. *Genomics* **45**: 244–9.

Moyle LC. 2008. Ecological and evolutionary genomics in the wild tomatoes (Solanum sect. Lycopersicon). *Evolution*. **62**: 2995-3013.

Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GLG, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J, et al. 2002. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296**: 1661–71.

Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, Gundlach H, Spannagl M. 2013. MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Res* **41**: D1144–51.

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. 2008. Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res* **18**: 2024–33.

Parkin I a P, Gulden SM, Sharpe AG, Lukens L, Trick M, Osborn TC, Lydiate DJ. 2005. Segmental structure of the Brassica napus genome based on comparative analysis with Arabidopsis thaliana. *Genetics* **171**: 765–81.

Pecinka A, Dinh HQ, Baubec T, Rosa M, Lettner N, Mittelsten Scheid O. 2010. Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in Arabidopsis. *Plant Cell* **22**: 3118–29.

Pecinka A, Schubert V, Meister A, Kreth G, Klatte M, Lysak M a, Fuchs J, Schubert I. 2004. Chromosome territory arrangement and homologous pairing in nuclei of Arabidopsis thaliana are predominantly random except for NOR-bearing chromosomes. *Chromosoma* **113**: 258–69.

Van de Peer Y, Fawcett J a, Proost S, Sterck L, Vandepoele K. 2009. The flowering world: a tale of duplications. *Trends Plant Sci* **14**: 680–8.

Pennacchio L a, Ahituv N, Moses AM, Prabhakar S, Nobrega M a, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**: 499–502.

Poitras L, Yu M, Lesage-Pelletier C, Macdonald RB, Gagné J-P, Hatch G, Kelly I, Hamilton SP, Rubenstein JLR, Poirier GG, et al. 2010. An SNP in an ultraconserved regulatory element affects Dlx5/Dlx6 regulation in the forebrain. *Development* **137**: 3089–97.

Poulin F, Nobrega M a, Plajzer-Frick I, Holt A, Afzal V, Rubin EM, Pennacchio L a. 2005. In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics* **85**: 774–81.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* **444**: 444–54.

Rosa S, De Lucia F, Mylne JS, Zhu D, Ohmido N, Pendle a., Kato N, Shaw P, Dean C. 2013. Physical clustering of FLC alleles during Polycomb-mediated epigenetic silencing in vernalization. *Genes Dev* **27**: 1845–1850.

Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE. 2002. Using the transcriptome to annotate the genome. *Nat Biotechnol* **20**: 508–12.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison CA, Slocombe PM, Smith M. 1977. Nucleotide sequence of bacteriophage φX174 DNA. *Nature* **265**: 687–95.

Schmitz RJ, Schultz MD, Urich M a, Nery JR, Pelizzola M, Libiger O, Alix A, McCosh RB, Chen H, Schork NJ, et al. 2013. Patterns of population epigenomic diversity. *Nature* **495**: 193–8.

Schoenfelder S, Sexton T, Chakalova L, Cope NF, Horton A, Andrews S, Kurukuti S, Mitchell J a, Umlauf D, Dimitrova DS, et al. 2010. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* **42**: 53–61.

Schranz ME, Song BH, Windsor AJ, Mitchell-Olds T. 2007. Comparative genomics in the Brassicaceae: a family-wide perspective. *Curr Opin Plant Biol*. **10**: 168-75.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–50.

Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, Steige K, Platts AE, Escobar JS, Newman LK, et al. 2013. The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat Genet* **45**: 831–5.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annu Rev Med.* **61**: 437-55.

Stein L. 2001. Genome annotation : from sequence to biology. *Nat Rev Genet* **2**: 493–503.

Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, et al. 2003. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biol* **1**: E45.

Stephen S, Pheasant M, Makunin I V, Mattick JS. 2008. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol Biol Evol* **25**: 402–8.

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, et al. 2008. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res* **36**: D1009–14.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. *Science (80-)* **320**: 486–8.

Tautz D. 2000. Evolution of transcriptional regulation. *Curr Opin Genet Dev* **10**: 575–579.

Tuskan G a, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov a, et al. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313**: 1596–604.

Ureta-Vidal A, Ettwiller L, Birney E. 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet* **4**: 251–62.

Varki A, Altheide TK. 2005. Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Res* **15**: 1746–58.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans C a, Holt R a, et al. 2001. The sequence of the human genome. *Science* **291**: 1304–51.

Visel A, Prabhakar S, Akiyama J a, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio L a. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40**: 158–60.

Viturawong T, Meissner F, Butter F, Mann M. 2013. A DNA-Centric Protein Interaction Map of Ultraconserved Elements Reveals Contribution of Transcription Factor Binding Hubs to Conservation. *Cell Rep* 1–15.

Wang J, Lee AP, Kodzius R, Brenner S, Venkatesh B. 2009. Large number of ultraconserved elements were already present in the jawed vertebrate ancestor. *Mol Biol Evol* **26**: 487–90.

Ware D, Jaiswal P, Ni J, Pan X, Chang K, Clark K, Teytelman L, Schmidt S, Zhao W, Cartinhour S, et al. 2002. Gramene: a resource for comparative grass genomics. *Nucleic Acids Res* **30**: 103–5.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–62.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**: 973–82.

Winkler H. 1920. *In Verbreitung und Ursache der Parthenogenesis im Pflanzen-und Tierreiche.* Gustav Fischer, Jena, Germany.

Womack JE. 2005. Advances in livestock genomics: opening the barn door. *Genome Res.* **15**: 1699-705.

Zheng BS, Yang L, Zhang WP, Mao CZ, Wu YR, Yi KK, Liu FY, Wu P. 2003. Mapping QTLs and candidate genes for rice root traits under different water-supply conditions and comparative analysis across three populations. *Theor Appl Genet.* **107**: 1505-15.

CHAPTER 2

Computational Analysis and Characterization of UCE-like Elements (ULEs) in Plant Genomes

Konstantinos Kritsas[1], Samuel E. Wuest[1], Daniel Hupalo[2], Andrew D. Kern[3], Thomas Wicker[1] and Ueli Grossniklaus[1]

[1] Institute of Plant Biology & Zürich-Basel Plant Science Center, University Zürich, Zollikerstrasse 107, 8008 Zürich, Switzerland

[2] Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire, USA

[3] Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA

Published in *Genome Research* **22**:2455-66

Abstract

Ultraconserved elements (UCEs), stretches of DNA that are identical between distantly related species, are enigmatic genomic features whose function is not well understood. First identified and characterized in mammals, UCEs have been proposed to play important roles in gene regulation, RNA processing, and maintaining genome integrity. However, all of these functions can tolerate some sequence variation, not explaining their ultraconserved and ultraselected nature. We investigated whether there are highly conserved DNA elements without genic function in distantly related plant genomes. We compared the genomes of *Arabidopsis thaliana* and *Vitis vinifera*; species that diverged ~115 Mya. We identified 36 highly conserved elements with at least 85% similarity that are longer than 55 bp. Interestingly, these elements exhibit properties similar to mammalian UCEs, such that we named them UCE-like Elements (ULEs). ULEs are located in intergenic or intronic regions and are depleted from segmental duplications. Like UCEs, ULEs are under strong purifying selection, suggesting a functional role for these elements. As their mammalian counterparts, ULEs show a sharp drop of A+T content at their borders and are enriched close to genes encoding transcription factors and genes involved in development, the latter showing preferential expression in undifferentiated tissues. By comparing the genomes of *Brachypodium distachyon* and *Oryza sativa*, species that diverged ~50 Mya, we identified a different set of ULEs with similar properties in monocots. The identification of ULEs in plant genomes offers new opportunities to study their possible roles in genome function, integrity, and regulation.

Introduction

An increasing number of studies indicates that although the larger part of eukaryotic genomes consists of non-protein coding DNA, this is far from being non-functional. Conserved non-coding sequences (CNSs) are found in large numbers in all animal genomes (Dermitzakis et al. 2002; Dermitzakis et al. 2004). CNSs are still conserved between humans and pufferfish, which diverged 450 Mya (Woolfe et al. 2005). Their average sequence identity varies depending on the genomes compared.

There are varying degrees of conservation of CNSs, with non-coding ultraconserved elements (ncUCEs) forming the extreme end of the distribution. UCEs were first identified as DNA stretches that are 100% identical between the mouse, rat, and human genomes over at least 200 bp (Bejerano et al. 2004). NcUCEs were mainly described among eutherian genomes, such as human, mouse, rat, dog, and cow (Bejerano et al. 2004; Stephen et al. 2008; Elgar, 2009). Although most ncUCEs only appeared during tetrapod evolution (Stephen et al., 2008), many were already present in the jawed vertebrate ancestor, spanning ~530 Mya of evolutionary time; however, their conservation falls off to ~80% (Wang et al. 2009). Because we currently do not know any biological process that would not tolerate at least some sequence variation, the function of these ultraconserved and ultraselected elements is enigmatic.

The majority of the ncUCEs and CNSs seem to be under purifying selection, indicating that they are not mutation cold spots but are strongly constrained functional elements (Drake et al. 2006; Chen et al. 2007; Katzman et al. 2007). In insects ncUCEs occur much less frequent and are smaller in size than the mammalian ones, thus being more similar to CNSs, which are often shorter and less conserved (Glazov et al. 2005).

In animals, ncUCEs and CNSs are enriched near specific functional groups of genes, e.g. encoding transcription factors and developmental regulators (Bejerano et al. 2004; Vavouri et al. 2007; Glazov et al. 2005). It was demonstrated that ncUCEs and CNSs can function as enhancers controlling tissue-specific gene expression (McEwen et al. 2009; Visel et al. 2008; Paparidis et al. 2007; Pennacchio et al. 2006; Woolfe et al. 2005). Nevertheless, their role as enhancers is not sufficient to explain their high conservation, because all protein-DNA, DNA-DNA or DNA-RNA interactions known to date tolerate significant sequence divergence without affecting their functions (Ludwig et al. 2000; Romano and Wray 2003; Ludwig et al. 2005; Poulin et al. 2005; Rastegar et al. 2008). Therefore, ncUCEs and CNSs are likely to serve additional - so far unknown - functions

that constrain their sequence.

Because ncUCEs are often single copy sequences and strongly depleted from segmental duplications and human copy number variants (Derti et al. 2006; Chiang et al. 2008), it was suggested that they could serve as genome integrity retention agents that act in a copy counting mechanism for chromosomes (Derti et al. 2006). In other words, ncUCEs in diploid cells should be present in exactly two copies to ensure genome integrity. In order to accurately assess their number, ncUCEs would have to be identical in sequence to avoid interactions with duplicated genomic regions. However, sequence retention and extreme conservation does not mean that they are essential for viability. In fact, deletion of four ncUCEs in the mouse did not cause obvious phenotypic abnormalities (Ahituv et al. 2007). Nonetheless, mutations in ncUCEs are deleterious over evolutionary time as evidenced by the fact the ncUCEs are under stronger selection than protein-coding regions (Katzman et al. 2007).

Until now little is known about the occurrence of CNSs in plant genomes. Most plant CNSs described to date are relatively small and reside close to genes. In monocots, apart from three exceptions (Bossolini et al. 2007; Wicker et al. 2008), most CNSs are short (average 20bp), flanking a small number of orthologous genes (Kaplinsky et al. 2002; Guo and Moose 2003; Inada et al. 2003). A recent study describes the existence of long identical sequences (over 100bp) between plant genomes; however, the reported sequences are part of regions of known function or origin, such as repeats, exons, or organellar DNA (Reneker et al. 2012).

Here, we focus on the identification of large UCE-like elements (ULEs) in dicot and monocot genomes. Special care was taken to ensure that ULEs are not part of any genic sequence with known function. By comparing the genome sequences of *Arabidopsis thaliana* and *Vitis vinifera* (grapevine), we identified 36 large and highly conserved ULEs, which are over 55 bp long and share at least 85% sequence identity. The divergence time between the two species is estimated to be 115 Mya (Fawcett et al. 2009), allowing significant changes in DNA sequence to occur. Monocots have their own set of ULEs and many are shared by the more closely related genomes of *Brachypodium distachyon*, *Oryza sativa* (rice), *Sorghum bicolor* (sorghum), and *Zea mays* (maize). Strikingly, despite a complete lack of sequence similarity between plant ULEs and animal ncUCEs, they share common properties, indicating that the evolutionary conservation of ULEs and ncUCEs may result from similar functional constraints and selective pressures in plants and animals.

Results

Identification of plant UCE-like elements (ULEs) between the *Arabidopsis thaliana* and *Vitis vinifera* genomes

In order to identify ULEs in plants, whole-genome comparisons of the two dicot species *Arabidopsis thaliana* and *Vitis vinifera* were performed. Among dicots with sequenced genomes *Vitis* is the most distantly related to *Arabidopsis*. The genome of *Arabidopsis* was used as anchor for the ULE search against *Vitis*. We define a ULE as a non-coding DNA sequence sharing at least 85% identity. To exclude that these sequences serve as potential transcription factor binding sites, we searched the *Arabidopsis* Gene Regulatory Information Server (AGRIS), a data base for transcription factor binding sites (Palaniswamy et al. 2006). The average size of the 763,000 predicted *cis*-regulatory elements is 6.4 bp. Often, such transcription factor binding sites are clustered, leading to larger conserved stretches (Davidson, 2001). Using AGRIS, we found 28 large putative transcription factor binding sites or clusters (> 25 bp) with the biggest being 50 bp. Thus, we searched for ULEs that were longer than 55 bp.

The *Arabidopsis* genome was spilt in fragments of 1,200 bp with a 600 bp sliding window and 600 bp overlap. These fragments were used in BLASTN searches against the *Vitis* genome. All conserved sequences over 55 bp with ≥85% similarity were investigated further, using a set of stringent criteria for the identification of ULEs (Table1): To exclude gene sequence motifs that may still have been present in this dataset, candidate sequences were used in BLASTN searches against all *Arabidopsis* coding sequences. BLASTN searches were also carried out against collections of *Arabidopsis* tRNAs, ribosomal genes, and known ncRNAs. The remaining sequences were used in BLASTN searches against mitochondrial and chloroplast DNA. Transposable elements were excluded from our dataset. The remaining candidates were used in BlastX searches against the non-redundant NCBI protein database to identify and eliminate any further protein-coding sequences that might not have been annotated in *Arabidopsis*. Finally, we removed conserved sequences overlapping intron-exon junctions as they might be part of alternative splicing products or wrongly annotated exons. To ensure that only ULEs of low copy number remained in our dataset, candidates with >5 copies were removed.

Table 1. Criteria for ULE identification

ULEs are:	**ULEs are not:**
1. >55 bp long	1. Coding sequences
2. ≥85% identity	2. tRNA, rRNA, ncRNA
3. Low copy number (≤5)	3. mtDNA, chlDNA
	4. Transposable elements
	5. *E. coli* contamination
	6. Encoding a protein motif
	7. In intron–exon junctions

In total, 36 candidate ULEs between the *Arabidopsis* and *Vitis* genomes met our criteria (Supplemental Table S1). The resulting ULEs reside in intergenic or intronic regions and all occur as single copies in the genome. We identified two paralogous elements, ULE27 and ULE28, which are found in tandem on chromosome 2. These ULEs are within 300 bp with each other. ULE27 is 2 bp longer than ULE28 but otherwise 100% identical. In *Vitis* ULE25 is found in two tandem copies within 250 bp on chromosome 4. One of the *Vitis* copies is 12 bp longer than the other but the shared sequences are 100% identical.

The ULEs comprise a total of 2,396 bp. The longest one is 105 bp and sequence identity ranges from 85% to 98%, with an average of 87.7%. Twenty-two were found in intergenic regions and 14 in introns. All ULEs were screened against *Arabidopsis* ESTs and novel transcripts detected after exosome depletion (Chekanova et al. 2007). For 28/36 ULEs there was no evidence of transcription, while the remaining 8 were at least partially covered by transcripts. The distribution of ULEs along the five *Arabidopsis* chromosomes is shown in Figure 1.

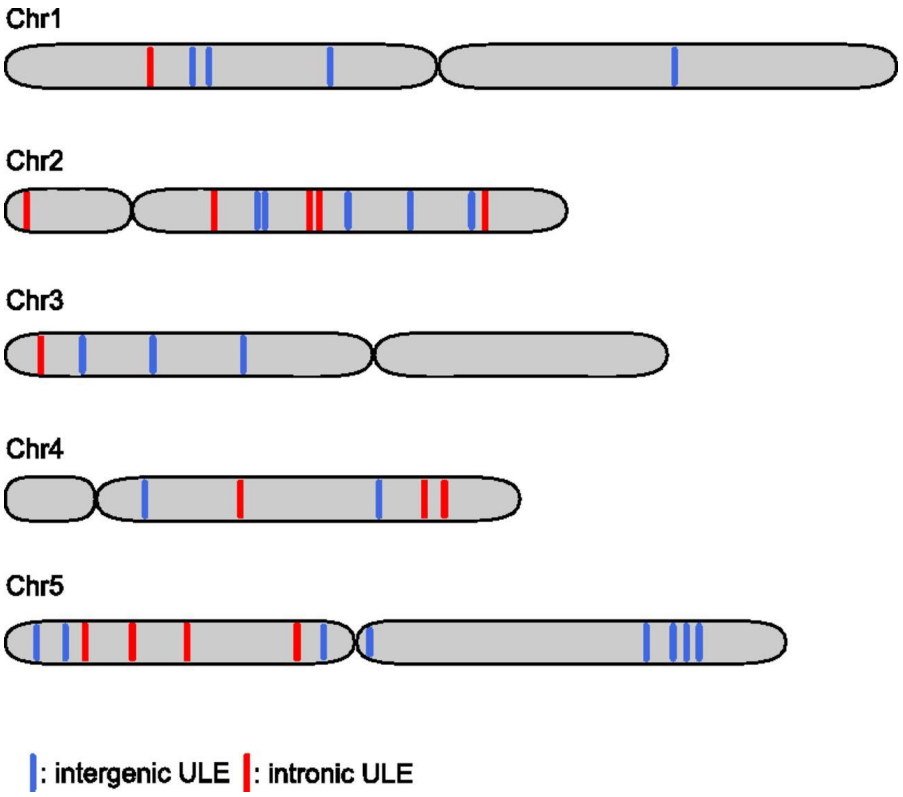


Figure 1. Distribution of ULEs along *Arabidopsis* chromosomes

Blue lines represent intergenic ULEs, red lines intronic ULEs. ULEs of both types are found on all chromosomes: On chromosome 1, ULEs are found on average every 6 Mb, while on chromosomes 2, 3, 4 and 5, ULEs are found in average every 1.9 Mb, 5.8 Mb, 3.8 Mb and 2.2 Mb, respectively.

ULEs are conserved among dicot but not more distantly related genomes

We tested whether the identified ULEs are present in other eudicot genomes, namely *Populus trichocarpa* (poplar), *Carica papaya* (papaya), *Cucumis sativus* (cucumber), and *Arabidopsis lyrata* (Tuskan et al. 2006; Ming et al. 2008; Huang et al. 2009) (Supplemental Table S2). The phylogenetic relationships of these species are shown in Figure 2. Twenty-two (22/36) were also present in the poplar genome, with similarities ranging from 83% to 98%, and a similar average identity as between *Arabidopsis* and *Vitis*. High levels of conservation were also found within the less complete genome of papaya, where 20 ULEs have identities ranging from 84% to 100%.

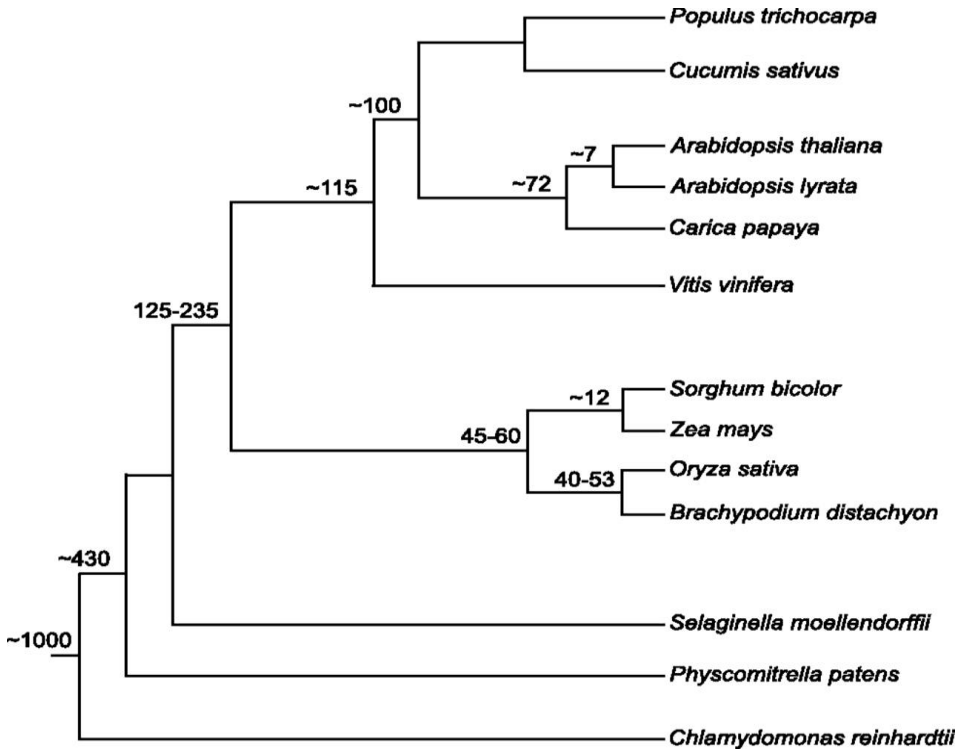


Figure 2. Phylogentic relatioships between major sequenced plant genomes

The phylogetic tree is adapted from phytozome.org. Divergence distances in Mya are indicated beside the nodes and taken from the following publications (Tuskan at al. 2006; Kuittinen et al. 2004; Chase et al. 2001; Fawcett et al. 2009; Davies et al. 2004; Yang et al. 1999; International Brachypodium Initiative 2010; Swigoňová et al. 2004; Stewart and Rothwell 1993; Yoon et al. 2004).

Only 9 of 36 ULEs were found in the cucumber genome with identities ranging from 85% to 98%. All but one of these corresponded to intronic ULEs, which suggests that scaffold data from cucumber are good enough for comparisons of genes but intergenic regions are not. Also, the genomes of poplar and papaya are less complete than the *Arabidopsis* genome, which may explain why not all ULEs were found. To test this, we investigated whether genes neighboring the ULEs that are not present in poplar or/and papaya, are also absent from those genomes. Indeed, the closest gene to these ULEs was not found or only partially present (less than a third of the corresponding sequence) in either the poplar or papaya genome (Supplemental Table S3). Finally, we looked for ULEs in the sequenced genome of another member of the mustard family, *A. lyrata* (Hu et al. 2011), where all but one ULE were conserved with identities between 93% and 100%.

We also searched for the 36 ULEs in the genomes of rice but found only one (ULE3) with 89% identity. ULE3 was partially conserved in two other monocot genomes, *Brachypodium* and maize (Supplemental Table S2). ULE3 is located upstream of gene *At2g33440*, which encodes an RNA-binding domain and is expressed at different developmental stages, but is functionally uncharacterized. None of the remaining ULEs, except for ULE19 in *B. distachyon*, were conserved. The identified ULEs were also searched against the genomes of the moss *Physcomitrella patens* and the green alga *Chlamydomonas reinhardtii* but no ULEs were conserved.

ULEs are mostly found in conserved collinear positions

In order to examine genes and other genic features that neighbor ULEs in *Arabidopsis* and *Vitis*, a genomic region spanning 3 kb from the 5’ and 3’ end of each ULE was analyzed. These 6 kb windows were used in BLASTN searches against the coding sequences of the two genomes. Among the 22 intergenic ULEs, 15 were located upstream of genes, three downstream of genes, and four in genomic regions where the nearest gene is more than 2 kb away.

To further assess ULE organization, we used the same 6 kb window and compared it by dotplot with an equivalent window in *Vitis*. To study whether ULEs are located in collinear regions, we classified flanking regions of ULEs as collinear when at least one of

the neighboring genes was homologous. We found that 29/36 ULEs were found in collinear regions (see example in Fig. 3A). Interestingly, 7 ULEs were found in non-collinear regions, where exclusively the ULE was conserved in the 6 kb segment, indicating that ULEs can be independent elements not necessarily associated with nearby genes (see example in Fig. 3B). For these 7 non-collinear ULEs we examined a larger region (50 kb) between *Arabidopsis* and *Vitis*: in 5 cases only the ULE was conserved (ULE2, ULE5, ULE9, ULE35, ULE36). It is intriguing that some of the ULEs are not found in collinear regions relative to *Vitis,* since in animals UCEs remain in collinear positions. One possible explanation for the non-collinear ULEs is that transposable element (TE) activity can lead to movement of genes and other sequences, thereby eroding collinearity (Wicker et al. 2010). Indeed, transposed genes in *Arabidopsis* are often associated with flanking repeats (Woodhouse et al., 2010), and this is also the case for three of the non-collinear ULEs (ULE2, ULE5, ULE35), which contain repeats within 3 kb of their borders (http://epigara.biologie.ens.fr/cgi-bin/gbrowse/a2e). Whether these repeats were associated with the movement of the ULEs or inserted afterwards cannot easily be distinguished.

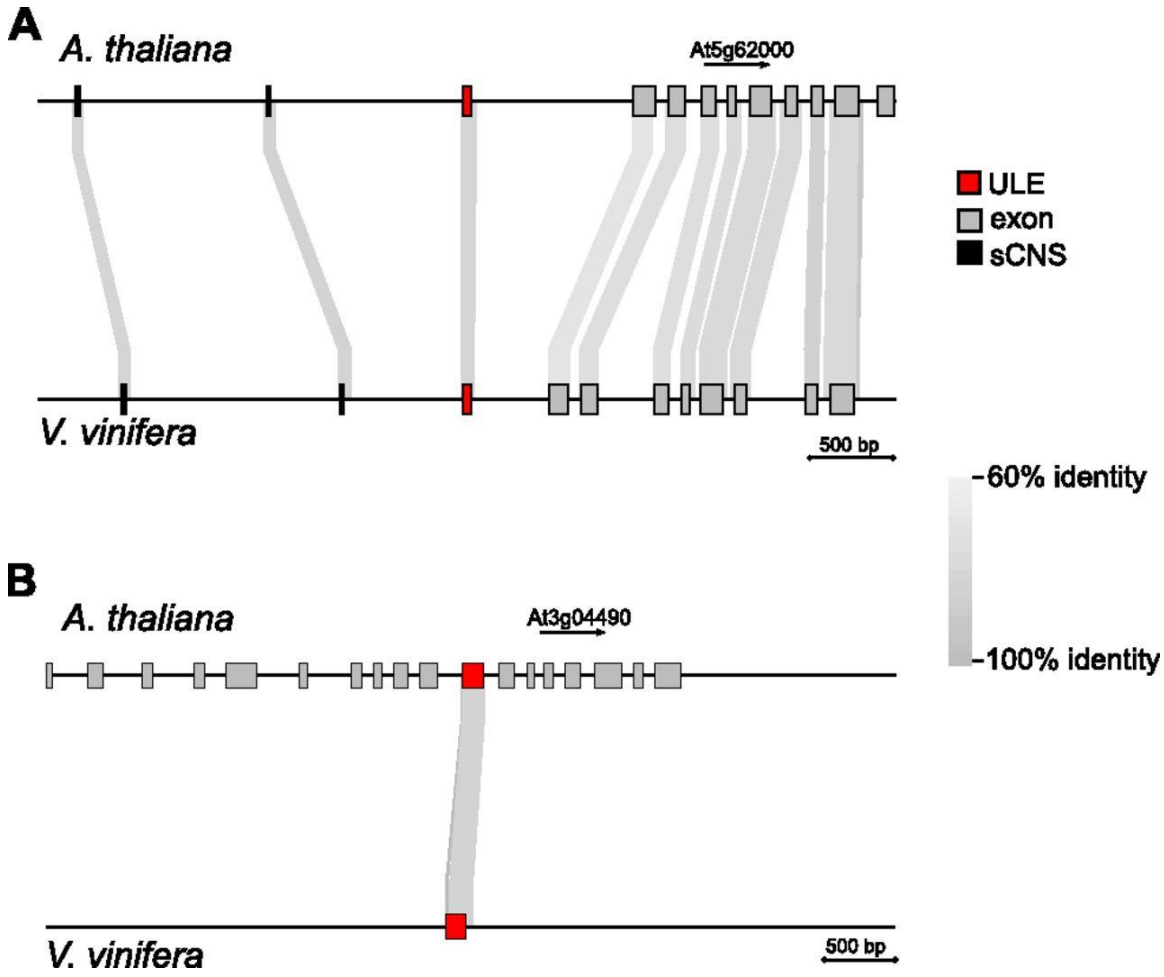


Figure 3. **Comparison of a 6 kb region surrounding two ULEs in *Arabidopsis* and *Vitis***

Conserved regions are indicated by shaded areas. ULEs are depicted in red, small conserved non-coding sequences (sCNSs) below 30bp in black, exons in grey.

(*A*) Comparison in collinear regions between *Arabidopsis* and *Vitis*.

(*B*) Comparison between non-collinear regions between *Arabidopsis* and *Vitis*.

ULEs are flanked by a sharp drop of the A+T content

To investigate whether ULEs have specific sequence characteristics, we compared the base composition at the boundaries of the ULEs, which are not conserved, with the one inside the ULEs (Fig. 4), as it was done for highly conserved non-coding sequences in vertebrates (Walter et al. 2005). We analyzed ULEs and their flanking regions in three blocks of sequences: 400 bp of flanking sequence plus 10 bp of the corresponding end of each ULE at the 5' and 3' border, and 30 bp from the middle of each ULE. We calculated the A+T content for each of these three blocks and observed a sharp drop in A+T frequency starting just before the borders of the ULEs. Within the ULEs, the A+T content was lower than in flanking regions (Fig. 4A). The same was observed when we analyzed the A+T frequency of each ULE individually (data not shown). We calculated the average A+T content in the *Arabidopsis* genome to be 63%, which is the same as the average A+T content in the regions flanking the ULEs (63%). In contrast, the average A+T content of the ULEs is 57% (Fig. 4A), and differs significantly from the A+T content of the sequences flanking the ULEs (0.57 *vs* 0.63, P=0.00104 by paired Wilcoxon signed-rank test). Thus, there is a sharp drop in A+T content at the borders of the *Arabidopsis* ULEs.

In *Vitis* we also observed a sharp drop of the A+T content at the ULE borders (Supplemental Fig. S1). The average A+T content of the *Vitis* genome is 65%, while the ULEs have an average A+T content of 57%. As in *Arabidopsis*, the A+T content of *Vitis* ULEs is significantly lower than that of the flanking sequences, which is 61% (0.57 *vs* 0.61, P=0.0103 by paired Wilcoxon signed-rank test).

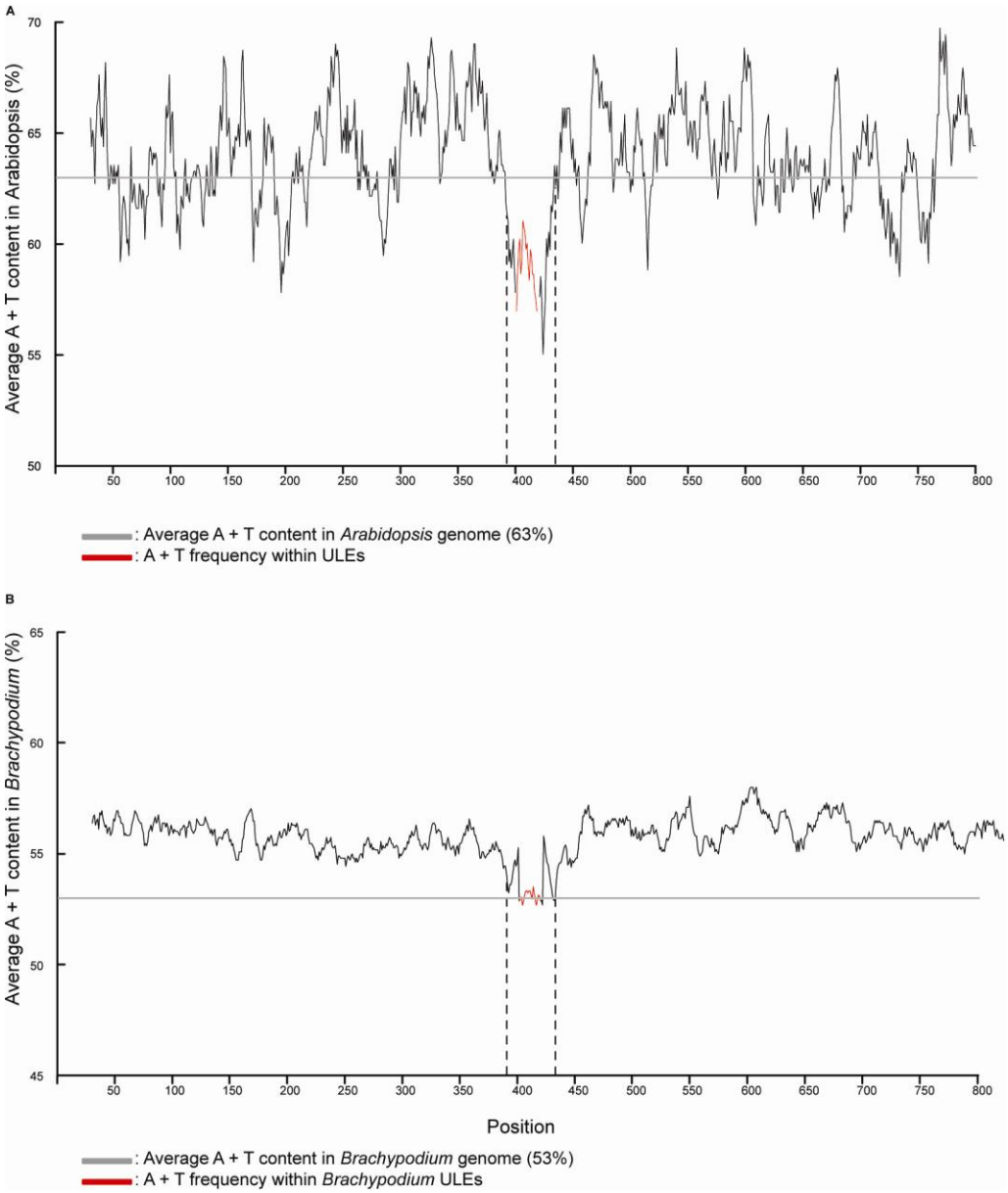


Figure 4. A+T content distribution within ULEs and their flanking regions

A+T frequency within ULEs is shown in red, whereas the frequency of flanking regions is shown black. The grey line depicts the average A+T content of the respective genome. Dashed vertical lines mark the last nucleotide of the neighbor regions before ULEs.

(A) A+T frequency in *Arabidopsis* ULEs (34/36).

(B) A+T frequency in *Brachypodium* ULEs (869/870) present in the genomes of rice, sorghum and maize, and their flanking regions.

ULEs are associated with specific functional categories of genes

We investigated whether ULEs are clustered near genes of distinct biological or molecular function. We examined the Gene Ontology (GO) annotations of genes flanking intergenic ULEs. For intronic ULEs, we considered only those genes in which they were located. ULE-flanking genes showed a significant enrichment for genes involved in development ($P \leq 2.2 \times 10^{-16}$). They also showed significant functional enrichment for genes associated with transcription factor activity ($P = 1.99 \times 10^{-5}$) and nucleic acid binding activity ($P = 3.5 \times 10^{-7}$) (Fig. 5).

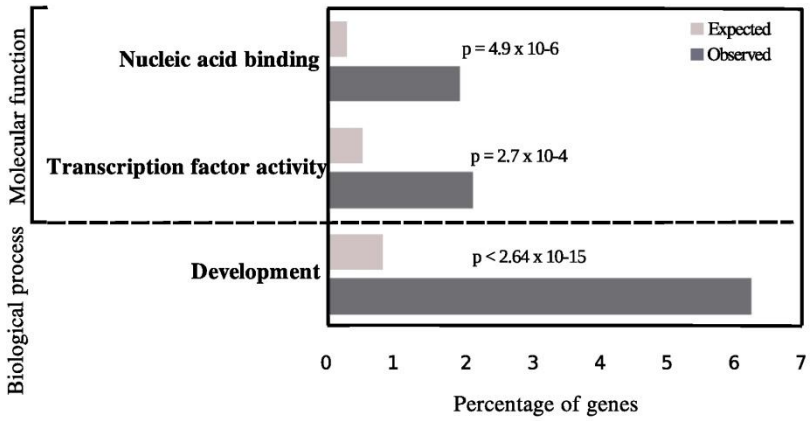
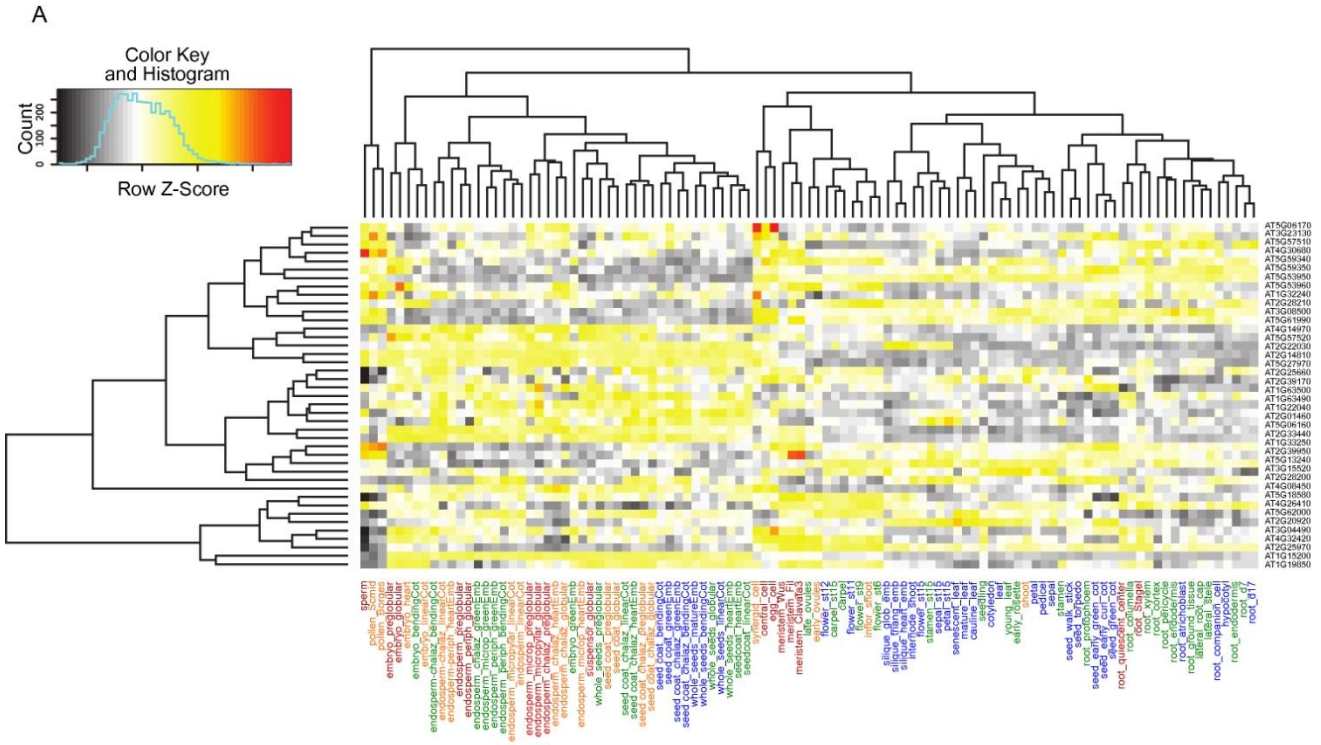


Figure 5. Expected versus observed percentage of genes in Gene Ontology annotation under the Molecular Function and Biological Process categories, corrected for multiple testing (Bonferroni correction).

ULE-associated genes exhibit expression peaks in undifferentiated tissues

Our GO analysis indicated that genes associated with ULEs are involved in development and are in turn likely to be developmentally regulated, too. In order to test this hypothesis, we estimated gene expression signals from a large collection of Affymetrix ATH1-array data querying a total of 103 different tissue and cell types of *Arabidopsis* (see Supplemental Table S4 and Methods for details). From 54 ULE-associated genes, 41 are targeted by probesets present on the ATH1-array. We visualized the average expression of ULE-associated genes across different developmental stages, tissues, and cell types. As shown in Figure 6A, several genes exhibit elevated expression

levels in gametophytes, embryo, and meristems. Figure 6B, summarizes the number of expression peaks found in distinct tissues of this developmental atlas. Because of small sample numbers, however, we could not test whether this increase is statistically significant. Furthermore, the dataset consists of a heterogeneous pool of data from different laboratories, tissue origins, and preparation protocols. Therefore, we classified tissues and cell types into four categories according to their differentiation state from (1) mainly undifferentiated cell populations to (4) mostly fully differentiated cell populations (Supplemental Table S4) and found that arrays from cell populations consisting of mainly undifferentiated cells (i.e. gametes, cells from the reproductive shoot meristem, early embryo and endosperm stages, as well as the root quiescent center) showed a significantly increased number of expression peaks (see Figure 6C; observed: 20, expected 12.4, P-value from randomly resampling 100,000 gene sets: P= 0.00939). From these results, we estimate that around 50% of ULE-associated genes show highest expression in an undifferentiated cell type. However, low expression of ULE-associated genes in other cell types does not necessarily exclude the importance of gene activity in these tissues. Overall, these results suggest that ULE-associated genes are developmentally regulated in plants and are often highly expressed in reproductive tissues.



B

Tissue Type	No. of expression peaks observed (expected)
meristem_Clavata3	4 (0.7)
egg_cell	3 (1.3)
sperm	3 (3)
embryo_preglobular	2 (0.5)
endosperm_chalaz_heartEmb	2 (0.8)
endosperm_micropylar_globular	2 (0.6)
meristem_FM	2 (0.3)
petal_st15	2 (0.5)
senescent_leaf	2 (0.8)
central_cell	1 (0.9)
embryo_globular	1 (0.6)
embryo_heart	1 (0.1)
endosperm_chalaz_globular	1 (1.3)
endosperm_chalaz_linearCot	1 (0.6)
endosperm_periph_bendingCot	1 (0.4)
endosperm_periph_globular	1 (0.7)
flower_st6	1 (0.2)
internode_shoot	1 (0.6)
lateral_root_cap	1 (0.3)
leaf	1 (0.3)
pollen_Borges	1 (1.2)
pollen_Schmid	1 (2.0)
root_endodermis	1 (0.2)
seedcoat_chalaz_heart	1 (0.2)
seedcoat_chalaz_linearCot	1 (0.3)
seedcoat_globular	1 (0.2)
silique_heart_emb	1 (0.2)
synergid_cell	1 (1.7)
Total:	41

Classification color key

1: cell population consisting primarily of undifferentiated

2: cell population containing undifferentiated cells or cells derived directly from undifferentiated cells

3: cell population containing some undifferentiated and some differentiated cells

4: cell population containing mainly fully differentiated cells

C

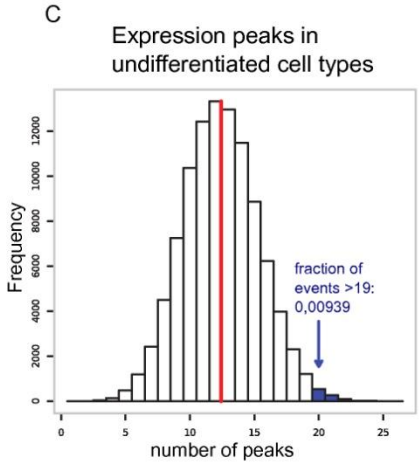


Figure 6. **ULE-associated genes are developmentally regulated**

(A) Heatmap representing color-coded relative expression among a large collection of *Arabidopsis* tissues/cell types. Dark colors indicate low expression and bright colors indicate high expression. Expression values were scaled per row (i.e. per gene) to visualize expression peaks of a transcript across developmental stages. Per-gene as well per-tissue, clustering was applied to visualize patterns in the

expression profiles. Sample descriptions are color coded as described in B. Only ULE-associated transcripts represented on the ATH1-array are shown (41/56).

(B) Table of tissues in which expression peaks of ULE-associated genes occur. The color-code indicates differentiation state of the respective tissue/cell type.

(C) ULE-associated gene expression peaks are significantly enriched in undifferentiated cells. The number of events where the maximal mean expression signal for one of 41 ULE-associated genes was found in undifferentiated cell types (i.e. gametes, shoot meristem cells, root quiescent center and early embryo/endosperm) is significantly higher than expected by chance. The histogram depicts the frequencies of expression peaks occurring in undifferentiated cell types amongst groups of 41 genes randomly sampled from the whole array. Resampling of random groups indicated that the same or higher number of expression peaks in undifferentiated cell types occurs only in 939 out of 100,000 instances (P=0.00939).

ULEs are depleted from segmental duplications

The fact that ULEs are single copy in the genome may indicate that multiple copies may be deleterious, possibly because that would interfere with the proposed copy counting mechanism (Derti et al., 2006). We searched whether ULEs are depleted from segmental duplications (SDs). During evolution, *Arabidopsis* has undergone multiple whole-genome and large-scale duplication events. We took into account SDs identified in *Arabidopsis* by Blanc and colleagues (Blanc et al., 2003), i.e. chromosome regions that share similar genes in the same order, excluding genes duplicated in tandem and transposable elements. In this survey 108 blocks of SDs sharing six or more duplicated genes were identified, which cover 71% of the *Arabidopsis* genome (80 Mb). The more recent duplications are estimated to have occurred 24-40 Mya (Blanc et al. 2003). Since these SDs refer to coding regions, we considered ULEs to be in segmental duplications when the closest gene to intergenic ULEs or genes containing intronic ULEs were within segmental duplications.

All intronic ULEs and, with the exception of one (At2g15510 flanking ULE8), all genes neighboring intergenic ULEs were outside SDs. To investigate the statistical significance of the identified trend for ULEs, a permutation test was applied in which 1,000 randomized datasets were sampled. Our test shows that the absence of ULEs from SDs is clearly non-random ($P<0.00036$). The depletion of ULEs from SDs indicates that they are dosage-sensitive and that there are selective constraints to keep them single copy.

ULEs are under purifying selection and not mutational cold spots

The high sequence conservation of ULEs between *Arabidopsis* and *Vitis* indicates that ULEs are selectively constrained sequences. Alternatively, their conservation could be due to the fact that they lie in regions with low mutation rates. In order to address this question, we estimated the distribution of selection coefficients from polymorphism data on the ULEs in 83 re-sequenced *Arabidopsis* accessions (Fig. 7) (Supplemental Table S5, Supplemental Table S6). The strength of selection acting on ULEs was compared relative to protein-coding regions and ULE-flanking regions (500 bp from the borders), respectively. Using the derived allele frequency (DAF) spectrum, we fit a Bayesian hierarchical model to estimate mean selection coefficients for each class of site. The hierarchical model was fit using a Markov Chain Monte Carlo while controlling for the effect of ascertainment on the ULE sites (Katzman et al. 2007; Kern 2009). The potency of removal of deleterious alleles increases as the selection coefficient decreases. Posterior estimates of mean selection coefficients between classes of sites indicates that ULEs may be under slightly stronger purifying selection than ULE flanking sites or exons, however as the credible sets overlap, such a difference is not statistically significant, only consistent with the hypothesis that ULEs might be under stronger purifying selection. This demonstrates that purifying selection rather than reduced mutation rates preserve ULEs at the DNA level. Thus, ULEs are under evolutionary pressure, which suggests that they are indeed functional elements.

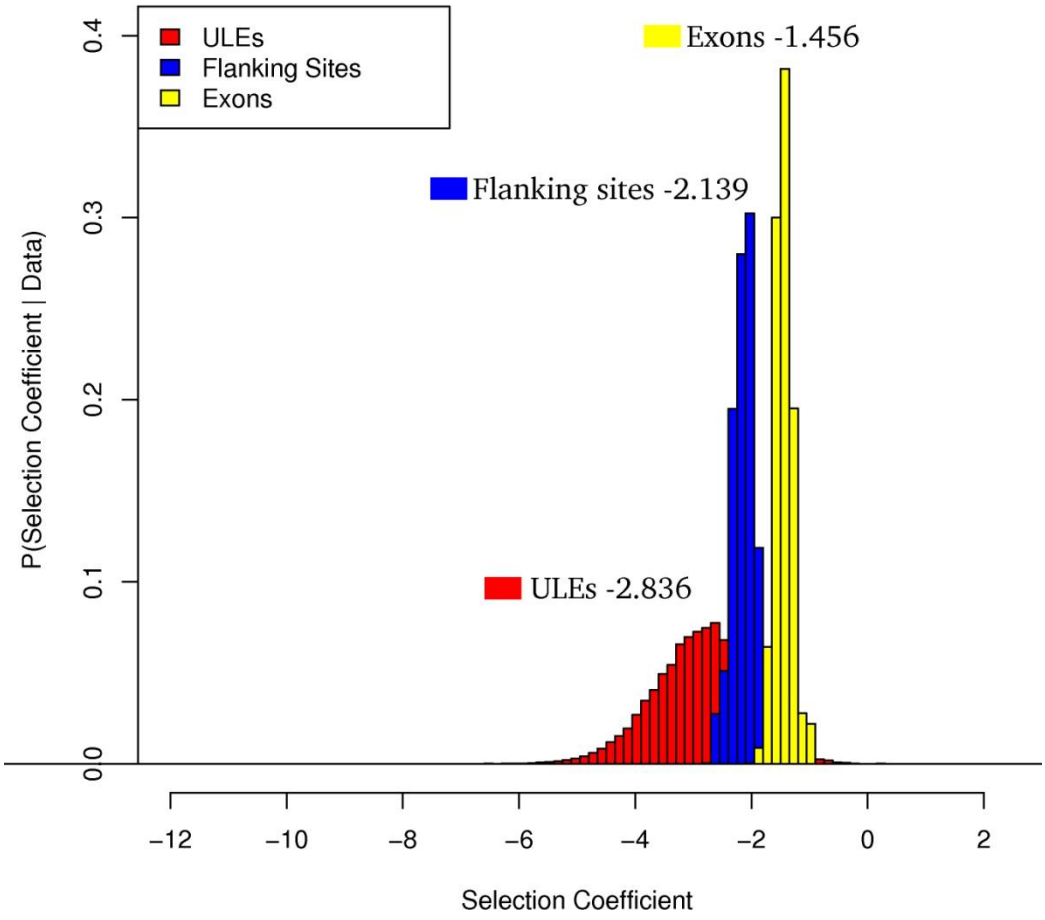


Figure 7. Selection coefficients for genomic regions

Shown are the posterior distributions of mean selection coefficients across classes of sites in the genome of *A. thaliana*. The values shown are α, the mean population scaled selection coefficient (2Nes). The values given are the MAP estimates from our MCMC (Supplemental Fig. S2, S3).

ULEs are not associated with recombination hotspots or origins of replication nor are they modified by DNA methylation

The observation that A+T content drops at the borders of the ULEs is intriguing because it implies a structural basis of these elements. Various cellular processes may be influenced by the A+T content, including recombination and replication. Indeed, it was shown that sequences with many ATs and TAs have lower recombination rates than those containing AGs, TCs, CAs, and TGs (Guo et al., 2009). Thus, we explored the possibility that ULEs are enriched at recombination hot spots (RHSs). RHSs are DNA regions with a higher rate of meiotic crossing-over than the surrounding DNA. In *Arabidopsis* studies in dense SNP regions from a sample of 19 accessions revealed around 260 RHSs, which

tend to occur in intergenic regions and are 1-2 kb long (Kim et al. 2007). None of the ULEs overlapped these RHSs. However, permutation test of 1,000 randomized datasets showed that the absence of ULEs from RHSs is not significant.

Further, we investigated whether ULEs are part of origins of DNA replication. Recently, ~1,500 putative origins of replication were mapped in *Arabidopsis* at a genome-wide scale (Costas et al. 2011). In this study, next generation sequencing was used to map newly synthesized DNA at the G1/S transition using synchronized cells. Only three ULEs are located within mapped origins of replication, the two tandem ULEs on chromosome 2 (ULE27, ULE28) and ULE33 on chromosome 1. However, this depletion of ULEs from origins of replication is not significant after applying a permutation test.

DNA methylation of cytosines is involved in epigenetic regulation. This epigenetic mark is heritably transmitted to following generations and affects various processes such as gene expression, genomic imprinting, transposon silencing, and timing of replication (reviewed in Vanyushin and Ashapkin, 2011). In *Arabidopsis*, the methylome at a single-base-pair resolution has been assessed in DNA of five week-old plants and flower buds, respectively (Cokus et al., 2008; Lister et al. 2008). The methylome of flower buds identified more than 2 million methylated cytosines accounting for 5.26% of genomic cytosines (Lister et al. 2008). We used this single-base-pair resolution DNA methylation map to investigate the methylation pattern of ULEs. The majority of the ULEs do not have any detectable methylation marks. Only seven ULEs are methylated in either the CG, CHH, CHG context (Supplemental Table S7). However, after applying a permutation test, the lack of ULE methylation is not statistically significant.

A distinct set of ULEs is shared between monocot genomes

Surprisingly, *Arabidopsis* ULEs were not present in genomes that are more distantly related than those of dicots. Thus, we asked whether there is another set of ULEs found explicitly in monocot genomes. We compared the genome of *Brachypodium distachyon* against that of *Oryza sativa* sb *japonica*. Divergence time between the two species is estimated at 40-53 Mya (International Brachypodium Initiative 2010), which is less than between *Arabidopsis* and *Vitis* (~115 Mya). We applied the same criteria as before (Table 1) and found 4,572 *Brachypodium* ULEs that are at least 85% identical to

rice and over 55 bp long. Median size and identity of these sequences is 69 bp and 87% respectively, similar to the ones found in dicots. Like the *Arabidopsis* ULEs, the majority of *Brachypodium* ULEs are single copy in the genome (4,491 out of 4,572). Interestingly, 870 sequences are also shared in the maize and sorghum genomes, which reflect conservation over 50 Mya (Fig. 2) (Supplemental Table S8) (International Brachypodium Initiative 2010).

We tested whether, apart from being single copy, the *Brachypodium* ULEs share other properties with *Arabidopsis* ULEs. Similarly, we calculated the A+T composition of these sequences relative to their flanking regions, which show no conservation (Fig. 4B). The average A+T content of the 870 *Brachypodium* ULEs shared with other monocots is 53%, which is identical to the average A+T content of the *Brachypodium* genome (53%), but differs significantly from that of the sequences flanking the ULEs, which is 55% (0.53 *vs* 0.55, P=0.000351 by paired Wilcoxon signed-rank test). Surprisingly, a similar drop of A+T composition at the borders of ULEs is present in both dicots and monocots. It is clear that *Brachypodium* and *Arabidopsis* ULEs, although distinct in sequence, share common characteristics.

Human UCEs are more abundant than *Arabidopsis* ULEs even when filtered under stricter criteria

Our results indicate that in plants ULEs are less common than UCEs are in mammalian genomes. However, in our study we used filter criteria that were more stringent compared to those used in mammalian studies. Thus, there might be fewer mammalian UCEs had they been analyzed under our criteria. To address this question, we reanalyzed the 481 UCEs identified by Bejerano and colleagues (2004). In our analysis we excluded UCEs within protein-coding sequences and functional ncRNAs and removed mitochondrial or *E. coli* sequences. In total, 390 elements, of 100% identity and length $\geq$ 200bp, meet the criteria we used, indicating that, even under these stringent criteria, mammalian ncUCEs are more abundant than plant ULEs.

Discussion

In this study we sought to identify and characterize highly conserved, non-coding elements in plant genomes. Synteny studies in flowering plant genomes revealed that the *Arabidopsis* genome is the most reshuffled, whereas the grapevine and papaya genomes have a better conserved ancestral genome structure (Huang et al. 2009). Thus, any conserved sequence between *Vitis* and *Arabidopsis* suggests a functional role. We focused on long stretches of conserved DNA (>55 bp), not necessarily associated with genes, in an unbiased search. In addition, our study used particularly stringent criteria in order to avoid any overlap with known genic sequences. Moreover, we were only interested in ULEs found at low copy number in the genome, thus targeting elements with a possible dosage effect that is prohibitive to accumulating high copy numbers.

Plant ULEs are fewer and less conserved than mammalian UCEs

One striking result from our comparative studies is the relatively low number of ULEs found in dicot genomes compared to sets of UCEs reported in mammals. If only the 390 mammalian ncUCEs that passed our filtering criteria are considered, the frequencies of these elements lie in the same range, i.e. one ncUCE/ULE per 8.0 Mb, 3.3Mb and 13.5 Mb in the human, *Arabidopsis* and *Vitis* genome, respectively. However, the ncUCEs identified by Bejerano and colleagues (2004) represent only the tip of the iceberg of the total number of highly conserved sequences. There are several thousands of UCEs (13'736) at least 100 bp long, which are shared between human and placental mammals (Stephen et al. 2008). In addition, there is a large number of conserved, non-coding elements that are slightly less than 100% identical (Dermitzakis et al. 2002; Woolfe et al. 2005). Thus, it appears that conserved non-coding sequences are more abundant in animals than in plants, perhaps because in animal genomes gene order is retained over millions of years (Li et al. 2010). More ULEs are lying in plant genomes when the genome comparison is made among less evolutionary distant plant species, such as the 870 we found shared by monocot genomes, with frequencies of one ULE per 0.4 Mb, 0.5 Mb, 0.9 Mb and 2.6 Mb in *Brachypodium*, rice, sorghum and maize, respectively. In fact, in contrast to dicot plants, monocots show a substantial conservation of gene order (International Brachypodium Initiative 2010).

The vast majority of the identified *Arabidopsis* ULEs arose after the divergence of dicots and monocots. However, *Arabidopsis* ULEs are well conserved in other dicot genomes, such as those of poplar, papaya, cucumber and *A. lyrata*, but between monocots and dicots only one ULE was retained. This is in sharp contrast to mammalian UCEs, where a major proportion covers an evolutionary time of ~530 Mya (Wang et al. 2009).

Why do plant genomes appear to contain fewer ULEs? One reason could be that plants and vertebrates have molecular clocks running at different speeds. It has been suggested that *Arabidopsis* has a faster molecular clock relative to other angiosperms (Paterson et al. 2010), whereas amniote evolution was accompanied by a slowdown in the molecular clock (Stephen et al. 2008). Thus, it is possible that plant ULEs evolved at a higher rate due to a faster molecular clock. This could also explain why ULEs are not conserved between monocots and dicots, as they might have diverged beyond recognition. Alternatively, our set of ULEs may represent distinct dicot and monocot innovations.

Plant genomes have the tendency to reorganize frequently, for example by undergoing whole-genome duplications (Masterson 1994). This could also contribute to the smaller number of ULEs, because genome duplication events might have relaxed the selective constraints on ULEs, allowing them to evolve faster.

ULEs from plants and animals have similar characteristics

A recent study identified a large number of highly conserved elements between sequenced plant and animal genomes but came to the conclusion that there are no sequences similar to mammalian UCEs in plants (Reneker et al. 2012). However, they did not filter out certain sequence classes, such as organellar DNA, rDNA, and *E. coli* contamination, as we did in our search for ULEs. Furthermore, their criteria were quite different from ours, and thus they could not identify the ULEs we report here. Although plant ULEs and mammalian ncUCEs are distinct sets of conserved sequences, they share a surprising number of common properties. Dicot ULEs and mammalian ncUCEs (Katzman et al. 2007) are under strong purifying selection. New alleles arising within ULEs may therefore be deleterious, making it unlikely that they become fixed in a population; hence their astounding sequence conservation.

We found that the A+T frequency is low at the borders of plant ULEs. The same feature is also shared among vertebrate and nematode conserved sequences (Walter et al. 2005;

Vavouri et al. 2007; Chiang et al. 2008). The fact that the drop in A+T content at the borders of ULEs and ncUCEs is a conserved feature between animal and plant genomes indicates that their function may have a structural basis. The A+T content can affect DNA topology, nucleosome positioning, and higher order chromatin organization (Segal et al. 2006; Hughes and Rando 2009), and influence DNA replication, repair, and recombination. However, ULEs do not appear to correlate with functional elements related to the structural features we tested and are not enriched in RHSs, origins of replication, or regions of DNA methylation. Like in ncUCEs from vertebrates and insects, the majority of dicot ULEs described in this study are found in the vicinity of genes involved in development and near genes whose molecular function is assigned to transcription factor activity. In addition, the majority of genes neighboring ULEs show strong expression in undifferentiated cells.

Based on the common properties between ULEs and mammalian ncUCEs, it is tempting to speculate that both sets of conserved sequences represent convergent evolutionary products that may be involved in the regulation of developmental genes. This is further supported by functional assays of ncUCEs showing that they act as enhancers during early embryo development in lamprey and mouse (Pennacchio et al. 2006; Visel et al. 2008). But why then are they so highly conserved? Enhancers usually do not require a high degree of sequence conservation (Stormo 2000) nor are they unusually large, even if clustered. Recent findings suggest that ncUCEs might have dual or even more functions, since part of the human ncUCEs are both transcribed and act as enhancers (Licastro et al. 2010). Except for enhancers, ULEs could potentially represent part of conserved *cis*-regulatory modules (CRMs), where one or more transcription factors bind to regulate the expression of neighbor genes. About 18,500 CRMs located upstream of genes are shared by *Arabidopsis* and poplar (Ding et al. 2012). Merely one (ULE6) out of thirteen intergenic ULEs tested is part of a such a CRM. In addition, in vertebrates a proportion of conserved non-coding elements (ncUCEs and CNSs) do not share common target genes in all six genomes tested (Sun et al. 2008). This finding suggests that mere *cis*-regulatory activity is unlikely the only explanation for the existence and high conservation of these elements.

Strikingly, ULEs are depleted from SDs in *Arabidopsis*, similarly to what was reported for mammalian ncUCEs (Derti et al. 2006). However, the existence of ULEs predates the existence of the segmental duplications we investigated in our analyses. This advocates that an evolutionary force kept the ULEs as single copies even though segmental duplications cover more than 70% of the *Arabidopsis* genome (Blanc et al. 2003). These

observations suggest that either ULEs *per se*, or the genomic regions that contain them, are dosage sensitive, and that a deviation from single copy could have an impact on the plant's fitness. These results also support the idea that ULEs function as agents involved in a chromosome copy counting mechanism (Derti et al. 2006). Here the maternal and paternal copies of ULEs/ncUCEs may recognize each other, perhaps through pairing, in order to determine the exact copy number of chromosomes, which in a diploid cell should be exactly two. Deviation from ULE/ncUCE copy number or sequence could trigger events that are deleterious to a cell with an abnormal number of chromosomes, but deleterious effects could also occur at the organismal or population level.

Despite the recent efforts to elucidate the function of conserved non-coding sequences, their role still remains elusive. ULEs have distinct characteristics and our data suggest that, in addition to sequence constraints, they are functional elements that are under purifying selection. Future studies are needed to shed light onto the purpose of their existence and their function.

Methods

Sequence Analyses

All analyses were performed on LINUX systems. For the identification of ULEs we developed software designed in PERL; all scripts are available upon request. Stand-alone BLAST software was obtained from NCBI (ncbi.nlm.nih.gov). For genome comparison studies, local BLAST databases were created. The *A. thaliana* genome sequence was downloaded from The Institute of Genomic Research (TIGR), now available from TAIR (arabidopsis.org). The genome of grapevine was obtained from Genoscope version1 (genoscope.cns.fr/externe/GenomeBrowser/Vitis), poplar version 1.1, *P. patens* version 1.1, and *C. reinhardtii* version 4.0 from the Joint Genome Institute (JGI) (jgi.doe.gov), *Oryza sativa* sb. *japonica* (rice) version 6 (rice.plantbiology.msu.edu/), *B. distachyon* from (brachypodium.org), maize version 2 from (plantGDB.org), papaya version 4 from (phytozome.net/papaya.php), and cucumber scaffold data from (cucumber.genomics.org.cn/page/cucumber/index.jsp). The sequence data from *A. lyrata* were produced by JGI in collaboration with the user community.

Coding, mitochondrial, and chloroplast sequences of *A. thaliana* were obtained from arabidopsis.org, TAIR9, ncRNA sequences from NCBI (ncbi.nlm.nih.gov, 08/1/2010) and PMRD: Plant microRNA database (Zhang et al. 2010). Transposable elements from the TREP database (wheat.pw.usda.gov/ITMI/Repeats), as well as repeats from the Plant Repeat Databases (plantrepeats.plantbiology.msu.edu/index.html). To identify possible *E. coli* contaminations, candidates were used in BlastN searches against the *E. coli* genome version NC 000313. The number of conserved sequences that were culled after applying the above filters is shown on Supplemental Table S9A.

Similar filters were applied for the identification of monocot ULEs (Supplemental Table S9B). *Brachypodium dystachion* (version 1.0) was used in BLASTN searches against the *Oryza sativa* sb *japonica* genome (version 6.0). Databases from (brachypodium.org, phytozome.org, rice.plantbiology.msu.edu, plantgdb.org) were used to discard candidates showing similarity to coding sequences, proteins, chloroplast and mitochondrial DNA. For repetitive elements PTREP and Plant Repeat Database were used. For small RNAs, PMRD, Cereal small RNAs database (http://sundarlab.ucdavis.edu/smrnas/), plant snoRNA database (http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna), and NCBI were used.

For annotation of the *Vitis* genes surrounding ULEs the two regions were aligned using

DOTTER (Sonnhammer and Durbin 1995) to determine positions of introns, exons, start and stop codons.

Characterization of ULEs

The A+T composition was calculated in a 10 bp window with a 1 bp sliding step width in each of the three sequence blocks. Two ULEs found in two closely spaced tandem copies were excluded from this analysis. For a comparison of A+T content of ULEs and flanking regions, ULEs were compared to sequences composed of one-half of the length of the ULEs flanking their 3' and 5' borders, respectively. A paired Wilcoxon signed-rank test was applied to assess significance.

For functional categorization, 54 genes located in flanking regions of intergenic ULEs or genes enclosing intronic ULEs were selected. Genes were grouped into different functional categories by using the TAIR9 Gene Ontology (GO) annotation. The data were compared with the functional categories assigned for all TAIR9 *Arabidopsis* genes. Fisher's exact tests were performed to determine over-representation of gene categories. P-values were corrected for multiple testing with the Bonferroni correction.

Expression analysis of ULE-associated genes

Original ATH1-array data from different *Arabidopsis* tissues were used as described (Wuest et al. 2010). Additional datasets were downloaded from public repositories (Supplemental Table S4). Data from the root quiescent center (Nawy et al. 2005), discrete seed compartments (Le et al. 2010), and cell types of the shoot apical meristem (Yadav et al. 2009) were added to the tissue atlas. The tissue data totally includes a set of 103 tissue types of gametophytic, sporophytic, and embryonic origin. Gene expression signals were calculated by dChIP (Version 2010) using invariant-set normalization and a PM-only model. Probeset definitions according a newer *Arabidopsis* genome release (TAIR9) were downloaded from (brainarray.mbni.med.umich.edu; ATH1-version 10, based on TAIR9 genomic sequences (Dai et al. 2005) and probes mapping to multiple probesets were removed from the analysis. For this, duplicated probe-sequences in the probeset definitions were identified in R (Version 2.8.1) and a

new chip description file generated using the Bioconductor package affxparser (Bengtsson et al. 2010, bioconductor.org). The mappings contain a total 21,253 probes mapping to unique gene identifiers (AGIs). From 56 ULE-associated genes, 41 were contained within the updated mappings. Log2-transformed dChip expression values were imported into R Version 2.11.1, where all subsequent analyses were performed. In order to simplify analyses, replicated array signals were averaged. Heatmaps were generated using functionality provided by the R-package gplots (Version 2.8.0) (Warnes et al. 2010).

Purifying selection of ULEs

The coordinates of the conserved elements were used to extract flanking regions that spanned 500 bp up- and downstream from each ULE. Genome annotation information from TAIR9 (arabidopsis.org) was used to randomly select a group of 50 coding sequences from the collection of all exons across the five *A. thaliana* chromosomes. 83 genomes, obtained from the ongoing 1001 *Arabidopsis* Genomes project (1001genomes.org; Cao et al. 2011), supplied variation data for the sequences in each group. Separately, sequences from all three groups of *A. thaliana* sequence were aligned to their *A. lyrata* and *V. vinifera* counterparts using Blast. The sequence at the node of the *A. lyrata/A .thaliana* phylogenetic precursor was ancestrally reconstructed using maximum likelihood as implemented in the PAML v4.3 software suite (Yang 2007) under a HKY85 nucleotide substitution model (Hasegawa et al. 1985). The ancestral sequence, aligned to the *A. thaliana* population data, provided a reference to determine whether the variations seen in the alignment were ancestral or derived. By parsing the collection of *A. thaliana* individuals and comparing variation to the ancestral sequence we were able to unfold a derived allele frequency (DAF) spectrum.

To estimate the strength of selection on each group of sequences, we took a hierarchical Bayesian approach. To fit our Bayesian hierarchical model we used the Markov Chain Monte Carlo (MCMC) algorithm described in Katzman et al (2007), which uses the Metropolis-Hastings algorithm for updates. Briefly, this model aims to estimate the mean and standard deviation of an unknown normal distribution representing the selective effect of new alleles in each of a series of "classes" of DNA (ULEs, exons, flanking sites). Individual alleles are each assumed to have their own selection coefficients, drawn

as independent, identically distributed random variables from this distribution. Further, selection coefficient estimates were corrected to account for divergence based ascertainment biases present in the ULE sequences (Kern 2009).

To evaluate the elements, flanking regions, and exonic regions, we ran 6 independent chains of 500,000 samples for the group of ULEs and for the flanking and exonic regions, respectively. To assess whether the chains converged, we plotted Gelman's potential scale reduction factor (Brooks and Gelman 1998) as implemented in the Coda R package (Plummer et al. 2006). After a reliable convergence, the first 25,000 iterations we discarded as burn in and the remaining samples we used to estimate a selection coefficient distribution as plotted in Supplemental Fig. S3.

Mammalian UCE analysis

Mammalian UCEs, the human genome (February 2009, hg19), and the mitochondrial genome were obtained from the University of California Santa Cruz (genome.ucsc.edu). UCEs were used in Blast searches against human cDNA sequences (enseble.org), eukaryotic tRNAs (gtrnadb.ucsc.edu), and ncRNAs from the Non-coding RNA database (biobases.ibch.poznan.pl/ncRNA). All mammalian UCEs that did not have matches in these datasets were used in BlastX searches against the non-redundant NCBI database to search for protein similarities. Subsequently, UCEs were used in BlastN searches against the *E. coli* genome.

Acknowledgements

We are indebted to Detlef Weigel, Jun Cao, Korbinian Schneeberger and Stephan Ossowski (Max Planck Institute for Developmental Biology, Tübingen) for providing access to SNP data for the ULEs in the *Arabidopsis* accessions they re-sequenced, and Sharon Kessler for comments in the manuscript. This work was supported by the University of Zürich and a Syngenta PhD-Fellowship of the Zürich-Basel Plant Science Center. DH and ADK are supported by Dartmouth College, the Neukom Institute, and NSF grant MCB-1052148.

References

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol* **5**: e234.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* **304**: 1321-25.

Bengtsson H, Bullard J, Gentleman R, Hansen KD, Morgan M. 2010. affxparser: Affymetrix file parsing. SDK. R package version 1.22.0

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res* **13**: 137-44.

Bossolini E, Wicker T, Knobel PA, Keller B. 2007. Comparison of orthologous loci from small grass genomes *Brachypodium* and rice: implications for wheat genomics and grass genome annotation. *Plant J* **49**: 704–17.

Brooks SP, Gelman A. 1998. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7:** 434-55.

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koening D, Lanz C, Stegle O, Lippert C et al. 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* **43**: 956-63.

Chekanova JA, Gregory BD, Reverdatto SV, Chen H, Kumar R, Hooker T, Yazaki J, Li P, Skiba N, Peng Q, et al. 2007. Genome-wide high-resolution mapping of exosome substrates reveals hidden features in the *Arabidopsis* transcriptome. *Cell* **131**: 1340-53.

Chen CT, Wang JC, Cohen BA. 2007. The strength of selection on ultraconserved elements in the human genome. *Am. J. Hum. Genet* **80:** 692-704.

Chiang CW, Derti A, Schwartz D, Chou MF, Hirschhorn JN, Wu CT. 2008. Ultraconserved elements: analyses of dosage sensitivity, motifs and boundaries. *Genetics* **180:** 2277-93.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452:** 215-19.

Costas C, Sanchez MP, Stroud H, Yu Y, Oliveros JC, Feng S, Benguria A, Lopez-Vidriero I, Zhang Z, Solano R et al. 2011. Genome-wide mapping of *Arabidopsis thaliana* origins of DNA replication and their associated epigenetic marks. *Nat Struct Mol Biol* **18:** 395-400.

Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H et al. 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* **33**: e175.

Davidson EH: *Genomic regulatory systems San Diego: Academic Press*; 2001.

Davies TJ, Barraclough TG, Chase MW, Soltis PS, Soltis DE, Savolainen V. 2004. Darwin's abominable mystery: Insights from a supertree of the angiosperms. *Proc Natl Acad Sci* **101**: 1904-09.

Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV, Antonarakis SE. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420:** 578-82.

Dermitzakis ET, Kirkness E, Schwarz S, Birney E, Reymond A, Antonarakis SE. 2004. Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res* **14:** 852-59.

Derti A, Roth FP, Church GM, Wu CT. 2006. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet* **38**: 1216-20.

Ding J, Hu H, Li X. 2012. Thousands of *cis*-regulatory sequence combinations are shared by *Arabidopsis* and poplar. *Plant Physiol* **158**: 145-55.

Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, et al. 2006. Conserved non-coding sequences are selectively constrained and not mutation cold spots. *Nat Genet* **38**: 223-27.

Elgar G. 2009. Pan-vertebrate conserved non-coding sequences associated with developmental regulation. *Brief Funct Genomic Proteomic* **8**: 256-65.

Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci* **106**: 5737-42.

Gaut Bs. 2002. Evolutionary dynamics of grass genomes. *New Phytol* **154:** 15-28.

Glazov EA, Pheasant M, McGraw EA, Bejerano G, Mattick JS. 2005. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of *homothorax* mRNA splicing. *Genome Res* **15**: 800-08.

Guo WJ, Ling J, Li P. 2009. Consensus features of microsatellite distribution: microsatellite contents are universally correlated with recombination rates and are preferentially depressed by centromeres in multicellular eukaryotic genomes. *Genomics* **93**: 323-31.

Guo H, Moose SP. 2003. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell* **15**: 1143-58.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**: 160-74.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet* **43**: 476-81.

Hughes A, Rando OJ. 2009. Chromatin 'programming' by sequence – is there more to the nucleosome code than %GC? *J Biol* **8**: 96.

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, et al. 2009. The genome of the cucumber, *Cucumis sativus* L. *Nat Genet* **41**: 1275-81.

Inada DC, Bashir A, Lee C, Thomas BC, Ko C, Goff SA, Freeling M. 2003. Conserved noncoding sequences in the grasses. *Genome Res* **13**: 2030-41.

International Brachypodium Initiative. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463:** 763-68.

Kaplinsky NJ, Braun DM, Penterman J, Goff SA, Freeling M. 2002. Utility and distribution of conserved non-coding sequences in the grasses. *Proc Natl Acad Sci* **99**: 6147-51.

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D. 2007. Human genome ultraconserved elements are ultraselected. *Science* **317**: 915.

Kern AD. 2009. Correcting the site frequency spectrum for divergence-based ascertainment. *PLoS One* **4**: e5152.

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D, Nordborg M. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat Genet* **39:** 1151-55.

Kuittinen H, de Haan AA, Vogl C, Oikarinen S, Leppälä J, Koch M, Mitchell-Olds T, Langley CH, Savolainen O. 2004. Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana*. *Genetics* **168**: 1575-84.

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, et al. 2010. The sequence and de novo assembly of the giant panda genome. *Nature* **463:** 311-317.

Le BH, Cheng C, Bui AQ, Wagmaister JA, Henry KF, Pelletier J, Kwong L, Belmonte M, Kirkbride R, Horvath S, et al. 2010. Global analysis of gene activity during *Arabidopsis* seed development and identification of seed-specific transcription factors. *Proc Natl Acad Sci* **107**: 8063-70.

Licastro D, Gennarino VA, Petrera F, Sanges R, Banfi S, Stupka E. 2010. Promiscuity of

enhancer, coding and non-coding transcription functions in ultraconserved elements. *BMC Genomics* **11**: 151.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133:** 523-36.

Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M. 2005. Functional evolution of a cis-regulatory module. *PLoS Biol.* **3:** e93.

Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*. **403:** 564-67.

Masterson J. 1994. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* **264**: 421-24.

McEwen GK, Goode DK, Parker HJ, Woolfe A, Callaway H, Elgar G. 2009. Early evolution of conserved regulatory sequences associated with development in vertebrates. *PLoS Genet* **5**: e1000762.

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* **452**: 991-96.

Nawy T, Lee JY, Colinas J, Wang JY, Thongrod SC, Malamy JE, Birnbaum K, Benfey PN. 2005. Transcriptional profile of the *Arabidopsis* root quiescent center. *Plant Cell* **17**: 1908-25.

Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E. 2006. AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol* **140**: 818-29.

Paparidis Z, Abbasi AA, Malik S, Goode DK, Callaway H, Elgar G, deGraaff E, Lopez-Rios J, Zeller R, Grzeschik KH. 2007. Ultraconserved non-coding sequence element controls a subset of spatiotemporal GLI3 expression. *Dev Growth Differ* **49**: 543-53.

Paterson AH, Freeling M, Tang H, Wang X. 2010. Insights from the comparison of plant genome sequences. *Annu Rev Plant Biol* **61**: 349-72.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444**: 499-502.

Plummer M, Best N, Cowles K, Vines K. CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6**: 7-11.

Poulin F, Nobrega MA, Plajzer-Frick I, Holt A, Afzal V, Rubin EM, Pennacchio LA. 2005. *In vivo* characterization of a vertebrate ultraconserved enhancer. *Genomics* **85**: 774-81.

Reneker J, Lyons E, Conant GC, Pires JC, Freeling M, Shyu CR, Korkin D. 2012. Long identical multispecies elements in plants and animal genomes. *Proc Natl Acad Sci* **109:** 1183-91.

Rastegar S, Hess I, Dickmeis T, Nicod JC, Ertzer R, Hadzhiev Y, Thies WG, Scherer G, Strähle U. 2008. The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. *Dev Biol* **318:** 366-77.

Romano LA, Wray GA. 2003. Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. *Devlopment* **130:** 4187-99.

Segal E, Fondufe-Mittendorf Y, Chen L, Thâström A, Field Y, Moore IK, Wang JP, Widom J. 2006. A genomic code for nucleosome positioning. *Nature* **442**: 772-78.

Sonnhammer EL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**: 1–10.

Stephen S, Pheasant M, Makunin IV, Mattick JS. 2008. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol Biol Evol* **25**: 402-08.

Stewart W, Rothwell GW. 1993. Paleobotany and the evolution of plants. 2nd edition. Cambridge University Press, Cambridge, UK.

Stormo GD. 2000. DNA binding sites: representation and discovery. *Bioinformatics* **16**: 16–23.

Sun H, Skogerbø G, Wang Z, Liu W, Li Y. 2008. Structural relationships between highly conserved elements and genes in vertebrate genomes. *PLoS One* **3:** e3727.

Swigoňová Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J. 2004. Close split of maize and sorghum genome progenitors. *Genome Res* **14:** 1916-23.

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313:** 1596-04.

Vanyushin BF, Ashapkin VV. 2011. DNA methylation in higher plants: past, present and future. Biochim Biophys Acta 1809: 360-68.

Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G. 2007. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biol* **8**: R15.

Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40**: 158-60.

Walter K, Abnizova I, Elgar G, Gilks WR. 2005. Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. *Trends Genet* **21**: 436-40.

Wang J, Lee AP, Kodzius R, Brenner S, Venkatesh B. 2009. Large number of ultraconserved elements were already present in the jawed vertebrate ancestor. *Mol Biol Evol* **26**: 487-90.

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T,

Maechler M, Magnusson A, Moeller S et al. 2010. gplots: Various R programming tools for plotting data. R package version 2.8.0

Wicker T, Narechania A, Sabot F, Stein J, Vu GT, Graner A, Ware D, Stein N. 2008. Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC Genomics* **9**: 518.

Wicker T, Buchmann JP, Keller B. 2010. Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res* **20:** 1229-37.

Woodhouse MR, Pedersen B, Freeling M. 2010. Transposed genes in *Arabidopsis* are often associated with flanking repeats. *PLoS Genet* **6**: e1000949.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**: e7.

Wuest SE, Vijverberg K, Schmidt A, Weiss M, Gheyselinck J, Lohr M, Wellmer F, Rahnenführer J, von Mering C, Grossniklaus U. 2010. *Arabidopsis* female gametophyte gene expression map reveals similarities between plant and animal gametes. *Curr Biol* **20**: 506-12.

Yadav RK, Girke T, Pasala S, Xie M, Reddy GV. 2009. Gene expression map of the *Arabidopsis* shoot apical meristem stem cell niche. *Proc Natl Acad Sci* **106**: 4941-46.

Yang YW, Lai KN, Tai PY, Li WH. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. *J Mol Evol* **48**: 597-604.

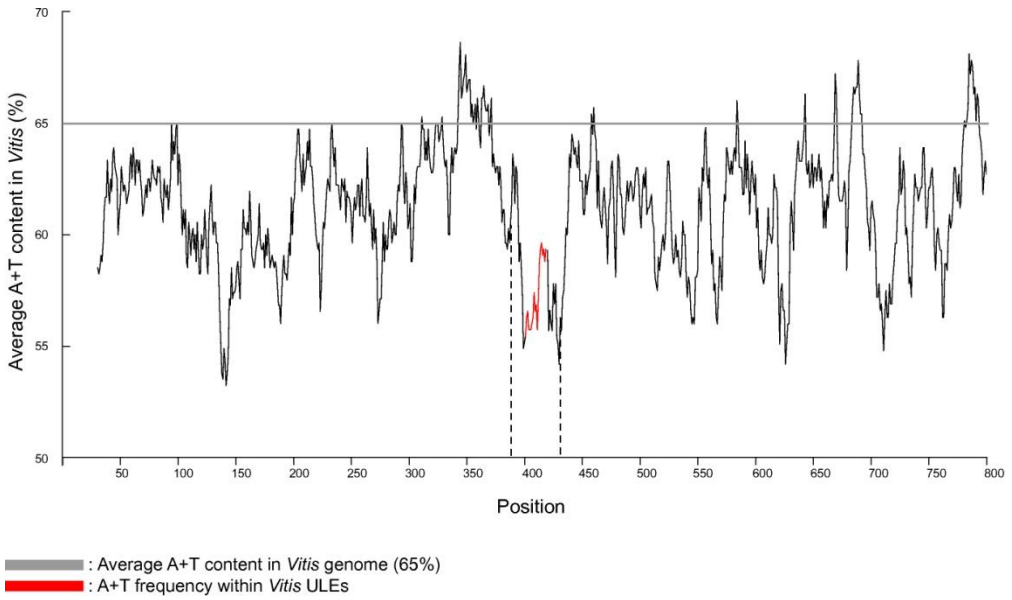
Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586-91.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol Biol Evol* **21**: 809-18.

Zhang Z, Yu J, Li D, Zhang Z, Liu F, Zhou X, Wang T, Ling Y, Su Z. 2010. PMRD: plant microRNA database. *Nucleic Acids Res* **38:** 806-13.

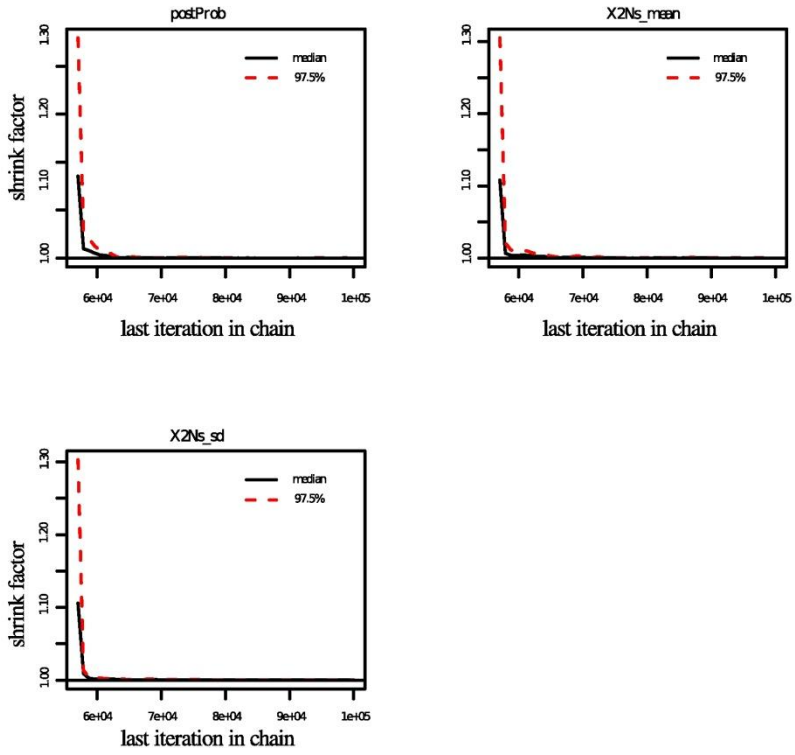
Supplementary materials

Supplementary Figures

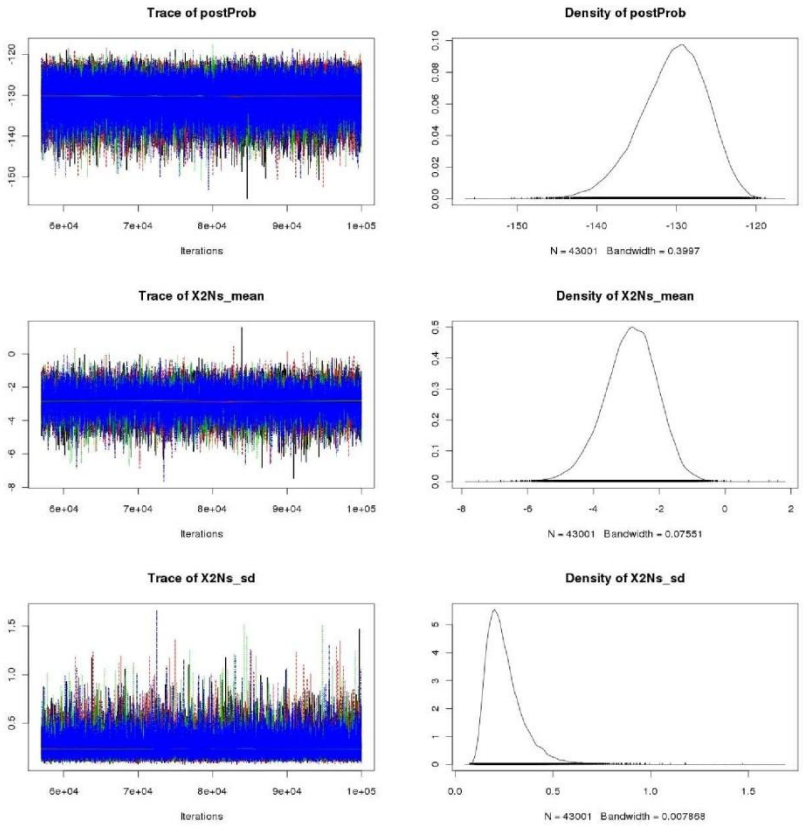


Supplementary Figure S1. A+T content distribution within ULEs and their flanking regions

A+T frequency within ULEs is shown in red, whereas the frequency of flanking regions is shown black. The grey line depicts the average A+T content of the respective genome. Dashed vertical lines mark the last nucleotide of the neighbor regions before ULEs. A+T frequency in *Vitis* ULEs (34/36).



Supplemental Figure S2. MCMC simulations were assessed using the scale reduction factor (Brooks Gelman 1998) for each set of chains. The progression of the random walk of the chains is plotted as iterations on the x-axis. If the independent chains begin to converge the value of the scale reduction factor, also known as the shrink factor, near 1.0.



Supplementary Figure S3. Plot of MCMC chains from one set of elements

On the left are six chains, each shown in different colors, starting from random independent points. The plots show three different views of the chains consisting of the joint posterior probability, μ, and σ, over the course of 500,000 iterations. The marginal distribution plot on the right helps visualize how well the chains converged.

CHAPTER 3

ULEs: novel functional elements hidden in the genome?

Konstantinos Kritsas[1], Celia Baroux[1] and Ueli Grossniklaus[1]

[1] Institute of Plant Biology & Zürich-Basel Plant Science Center, University Zürich, Zollikerstrasse 107, 8008 Zürich, Switzerland

Abstract

UCE-like elements (ULEs) are highly conserved non-coding sequences shared between plants. ULEs share common properties with non-coding UCEs (ncUCEs), the animal counterparts. It is believed that ULEs/ncUCEs retain their sequence identity due to essential functional properties. However, until today no satisfactory explanation justify their robust conservation. It has been suggested that the reason why ULEs/ncUCEs are depleted from segmental duplications is because they participate in a mechanism of chromosome copy counting through pairing and sequence comparison. Here, we explore this hypothesis. We provide evidence from fluorescence *in situ* hybridization experiments (FISH), that ULEs belong to chromosome regions of elevated somatic pairing frequency. In addition to somatic pairing, we further investigate the dosage nature of ULEs. T-NDA perturbation of one ULE yielded distorted transmission efficiency of the mutant in the offspring in a gender specific manner. However, transmission efficiency of the same mutant was normal in an aneuploid, trisomic background. No phenotypes were observed by addition of four extra copies of ULEs or by altering their methylation status.

Introduction

The advent of genome sequencing technologies has made it possible to identify DNA sequences of unknown function that potentially can contribute to the development organism. Genome sequence comparisons revealed that there is a large number of non-coding sequences which are evolutionary conserved and are present in animal and plant species (Dermitzakis et al. 2002; Siepel et al. 2005; Vavouri et al. 2007; Woolfe et al. 2005; Kritsas et al. 2012; Baxter et al. 2012; Haudry et al. 2013; Hupalo and Kern 2013). An extreme example of conservation are the non-coding ultraconserved elements (ncUCEs), stretches of DNA which are 100% identical between human, mouse and rat genomes and nearly as old as ~500 million years old (Bejerano et al. 2004; Glazov et al. 2005; Ovcharenko 2008; Stephen et al. 2008; Wang et al. 2009). Plants, as well have their own specific conserved non-coding sequences (CNSs) although they don't show the extreme conservation of ncUCEs, are fewer in number and smaller in size (Kritsas et al. 2012).

Previously, a set of UCE-like elements (ULEs) were identified after genome comparison studies between *Arabidopsis thaliana* and *Vitis vinifera* (grape), which diverged ~115 Mya, and between *Brachypodium distachyon* and *Oryza sativa* (rice), with a divergence time of ~50 Mya (Kritsas et al. 2012). Both sets of sequences (ncUCEs/ULEs) are clustered next to genes encoding transcription factors and genes involved in development (Bejerano et al. 2004; Kritsas et al. 2012). In addition, they both have a structural signature consisting of a sharp drop of A-T content at their borders exactly (Chiang et al. 2008; Kritsas et al. 2012). Moreover, both animal ncUCEs and plant ULEs are under negative selection suggesting that these elements are functionally important (Katzman et al. 2007; Kritsas et al. 2012). The striking commonalities between ncUCEs and ULEs make us speculate that both sets of sequences represent convergent evolutionary products which are invented independently to serve common functions.

But what is then the function of these elements? Although, there is a flood of genome comparison data and *in* silico characterization of CNSs our knowledge of their function still remains limited. For some ncUCEs and CNSs, it was shown that they act as tissue specific enhancers in *in vivo* assays using mice and zebrafish embryos (Woolfe et al. 2005; Poulin et al. 2005; Pennacchio et al. 2006; Visel et al. 2008). NcUCEs are also proposed to be involved in epigenetic regulation. In embryonic stem cells ncUCEs which coincide next to transcription factors are subject to bivalent chromatin modifications,

hence acquire both activating and repressing marks (Bernstein et al. 2006; Lee et al. 2006). Recent findings raise the possibility that ncUCEs in addition to enhancer activities could concurrently act at the transcriptional level (Licastro et al. 2010). Indeed, transcripts of a ncUCE can act as co-activators *in trans* to regulate homeodomain proteins (Feng et al. 2006) and another ncUCE is considered to affect the apoptosis of colon cancer cells *in vitro* (Calin et al. 2007). Recently, it has been proposed that ncUCEs are under evolutionary pressure because they are transcription factor binding hubs (Viturawong et al. 2013).

Based on the previous results, the current view for the role of (ncUCEs/ULEs) is that they behave as regulators of key developmental genes. In line with this, they are highly conserved because they have a significant role in the evolution of complex developmental programs. However, enhancer activity cannot fully explain the ultraconservation nature of ncUCEs/ULEs. It is known that enhancers do not require a high degree of conservation (Stormo 2000; Ghanem et al. 2003). In addition, ULEs are not part of *cis*-regulatory modules arguing against the concept of being a mosaic of regulatory elements (Kritsas et al. 2012).

Intriguingly, ncUCEs are depleted from segmental duplications and copy number variants (Derti et al. 2006; Chiang et al. 2008). Strikingly, ULEs are also depleted from *Arabidopsis* segmental duplications (Kritsas et al. 2012). In fact in both cases the existence of ncUCEs/ULEs predates the occurrence of segmental duplications suggesting that these elements or the regions that contain them are dosage sensitive. Thus, in addition to their regulatory role, they have been proposed to be involved in a chromosome copy counting mechanism (Derti et al. 2006). According to this model the two homologous chromosomes compare each other at the ncUCEs/ULEs regions and any deviations from the normal chromosome number could compromise the genome integrity. In accordance, deletions or additions of ncUCEs/ULEs may be deleterious for the organism or the population.

Here, in light of the chromosome copy counting model, we address whether plant ULEs are agents of such mechanism. We envision that the comparison of the homologous ULEs is resolved through pairing at their regions. Therefore, we carried out a fluorescence *in situ* hybridization (FISH) approach in somatic nuclei of *Arabidopsis thaliana*. We provide compelling evidence showing that the pairing frequency of homologous chromosomes is higher in ULE regions as oppose to non-ULE chromosome regions. We further assess the dosage sensitivity of the ULEs by taking advantage of insertional mutants. Perturbation of one ULE caused an increase of the transmission efficiency (TE)

of the mutant to the offspring. Interestingly, the phenotype disappears after just two generations. Segregation of the same mutant was not distorted in a trisomic background. No obvious phenotypes were observed by inserting four extra copies of ULEs and by altering their DNA methylation pattern.

Results

Pairing frequency of somatic homologous chromosomes is higher in ULE regions

If a chromosome copy counting mechanism exists, then prediction is that this mechanism should be mediated through chromosome pairing. Along this line, the two homologous chromosomes are getting closer, compare and recognize each other at the ULE regions. To address this question we employed a two-color fluorescence *in situ* hybridization (FISH) approach in the leaf somatic nuclei of *A. thaliana*.

Bacterial artificial chromosomes (BACs) that contain ULEs were labeled with digoxigenin (green color) and BACs right next to the BAC-ULEs with biotin (red color). Probes were hybridized against flow-sorted 2C rosette leave nuclei (Supplemental Figure 1). One signal per BACs pair was considered as homologous chromosomes pairing in that particular DNA region (Fig. 1A) whereas two separate signals indicated the absence of pairing (Fig. 1B). In some cases where the two signals were very close together (Fig. 1C); we regarded them as two separate signals.

First, we asked whether the observed frequencies of homologous pairing with our FISH settings are comparable with frequencies observed in previous FISH studies. We calculated the occurrence of a single signal of the regions R3 and R4. R3 and R4 are DNA regions on chromosome three and four respectively, and the pairing frequency has been assessed in Pecinka et al. 2004. In agreement, with that study, no statistical differences were detected with our observed pairing frequencies (Table 1).

Next, the pairing density of the ULE regions was assessed. Nine ULE regions and six regions that have no ULEs (non-ULE regions) were used in our study (Fig. 2). These regions span all five *Arabidopsis* chromosomes. The pairing frequency of each ULE region was then compared with the frequency of the non-ULE region lying on the same chromosome (Fig. 3) and statistical significance was calculated. For example, pairing frequency of ULE3 region is significantly different when compared with the R2 non-ULE region which is located on the same chromosome (Table 2).

Interestingly, the average frequency of a single signal from ULE regions is 20.8% whereas the corresponding frequency of non-ULE regions is 7.25%. Thus, on average positional pairing on ULEs is approximately 3-fold higher than other chromosome areas. Seven ULE-regions (ULE6, ULE3, ULE12, ULE4, ULE25, ULE11, and ULE1) showed

statistically significant higher pairing frequency when compared with the pairing frequency of the non-ULE regions (Table 2). This is true for ULEs that are located on all five chromosomes. Pairing occurs regardless of the genomic context surrounding ULEs since 5 of them (ULE1, ULE3, ULE6, ULE11, ULE12) lie in intergenic regions and two of them (ULE4, ULE25) in introns. Hence, in agreement with the chromosome copy counting hypothesis, our results reveal that somatic homologous pairing is significantly increased in ULE areas.

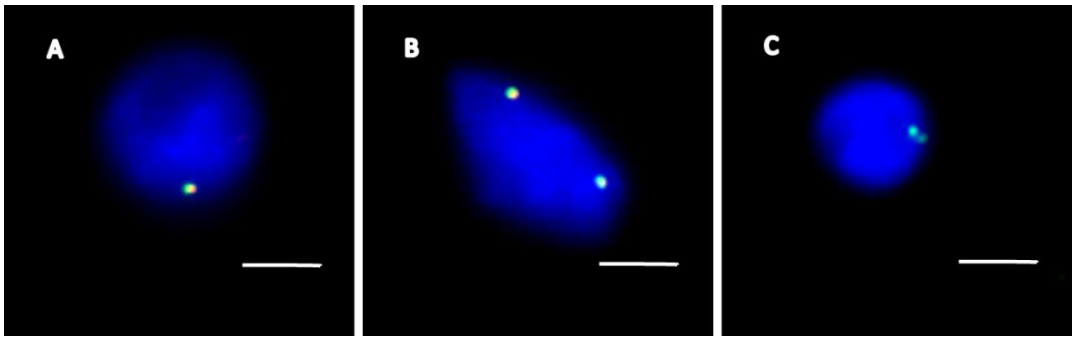


Figure 1. **Positional somatic homologous chromosome pairing in *Arabidopsis* 2C nuclei**

(A) Single point homologous pairing. *(B)* Unpaired homologous segments. *(C)* Homologous segments in close association less than the signal diameter.

In blue nuclei counterstained with DAPI. Green dot is BAC DNA labeled with digoxigenin, red dot adjacent BAC DNA labeled with biotin. Bar = 3 μm.

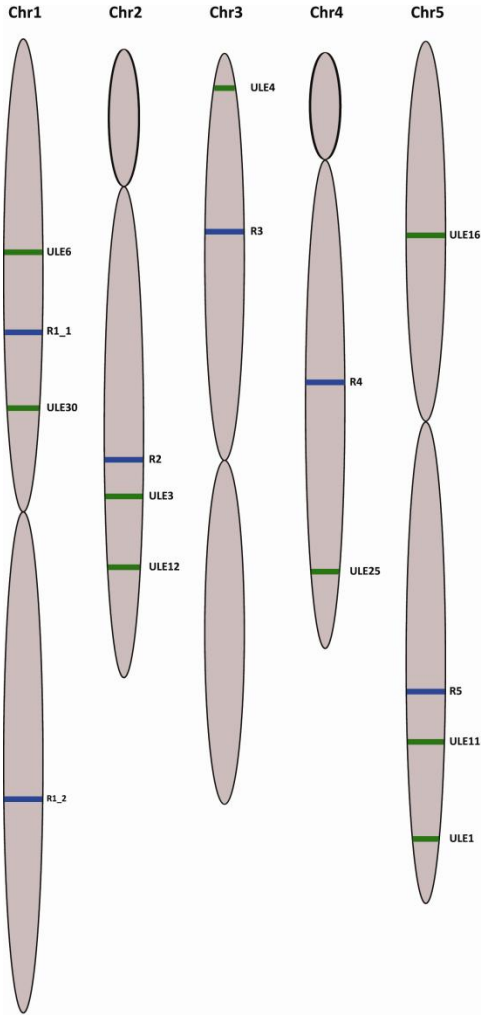


Figure 2. Distribution of BACs used for fluorescence *in situ* hybridization spanning the five *Arabidopsis* chromosomes.

Green color rectangular depict BAC region containing ULE, blue color rectangular depict non-ULE BAC region.

Table 1. Pairing frequency of homologous chromosome regions by fluorescence *in situ* hybridization and comparison with the pairing frequency of the same regions as it was published in Pecinca et al. 2004. Differences between the two pairing frequencies were tested with two-tailed Fisher's exact test.

BAC region	Chromosome	Nuclei	Pairing frequency	Nuclei Pecinca et al. 2004	Pairing frequency Pecinca et al. 2004	Two-tailed p-value
R3	3	106	7.54%	141	4.3%	0.28
R4	4	51	5.88%	107	3.7%	0.38

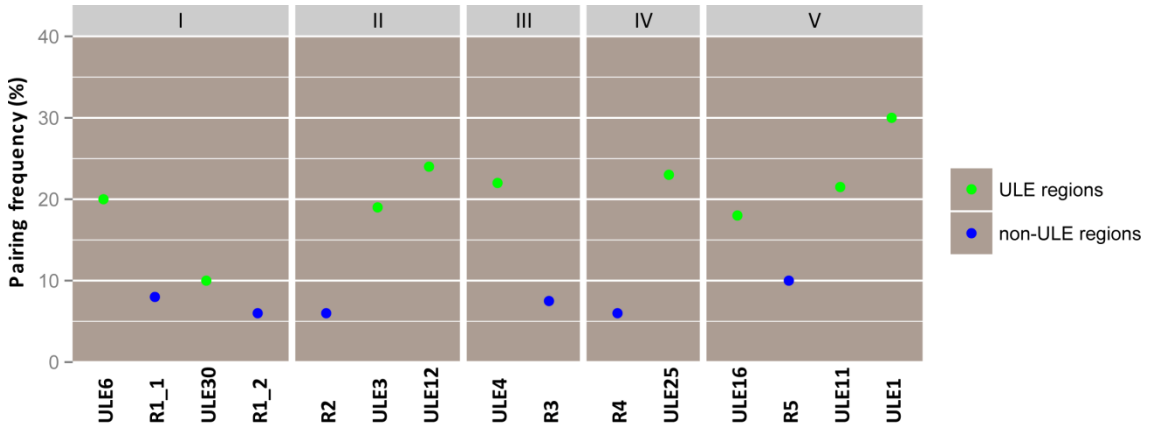


Figure 3. **Homologous chromosome pairing frequency in 2C *A. thaliana* rosette leaf nuclei**

On y-axis the frequency of single-point signal is indicated along the five *Arabidopsis* chromosomes. Green dots represent the pairing frequency of BAC-ULE regions and blue dots the pairing frequency of non-ULE BAC regions.

Table 2. Pairing frequency of homologous chromosome regions bearing ULEs by FISH and comparison with the pairing frequency of non-ULE chromosome regions. Two-tailed Fisher's exact test was applied to test the differences between pairing frequencies of a ULE-region with a non-ULE region on the same chromosome.

Chromosome	BAC region	Nuclei	Pairing frequency (%)	Two-tailed p-value
1	ULE6	76	19.73	0.02[1] 0.008[2]
1	ULE30	107	10.28	0.64[1], 0.31[2]
1	R1_1	110	8.18	-
1	R1_2	101	5.94	-
2	ULE3	108	24.07	$4x10^{-4}$
2	ULE12	105	19.04	0.01
2	R2	96	6.25	-
3	ULE4	108	22.22	$3x10^{-2}$
3	R3	106	7.54	-
4	ULE25	105	22.85	0.01
4	R4	51	5.88%	-
5	ULE16	110	18.18	0.12
5	ULE11	98	21.42	0.03
5	ULE1	83	30.12	$6x10^{-4}$
5	R5	108	10.18	-

[1] Fisher's test was applied between pairing frequency of ULE6, ULE30 and R1_1 regions
[2] Fisher's test was applied between pairing frequency of ULE6, ULE30 and R1_2 regions

Segregation of a chromosome carrying an insertion on ULE is distorted

ULEs were shown that they are not just single copy in the genome but are also selected to be absent from *Arabidopsis* segmental duplications (Kritsas et al. 2012). Hence, alteration of the ULE copy number may be a cause for reduced fitness. To address this, we searched for insertional T-DNA mutants that disrupt their sequence. We found three insertional mutants, one insertion each for ULE1, ULE6 and ULE30. All three ULEs are located in intergenic regions and are upstream of genes. ULE1 is ~1'150bp upstream of the Auxin Response Factor 2 (ARF2, At5g62000) gene; ULE6 is 700 bp upstream of the auxin-dependent transcription factor, Monopteros (At1g19850), and ULE30 ~280 bp upstream of the transcription factor Kanadi 2 (Kan2, At1g32240).

We then investigated whether these insertions confer severe developmental defects. For ULE1 and ULE30 insertions, no obvious phenotype was observed. Interestingly, homozygous *ule6/ule6* plants produced seedlings with phenotype similar to the *monopteros* mutant plants, the gene that is located downstream of ULE6 (Supplemental Figure 2). Mutations on *monopteros* gene eliminate the basal elements of the seedling such as hypocotyl, radicle and root meristem (Hardtke and Berleth 1998).

Although mutations on ULE1 and ULE30 did not cause any clearly visible phenotypes, we further investigated the TE of the mutation on each ULE, this allows to test the fitness of gametes carrying the disrupted ULEs. Southern blot hybridizations showed that both lines have insertions at a single locus (Supplemental Figure 3). We then used the hemizygous mutants in reciprocal crosses with wild type Col-0 plants (Table 3). *Ule1* did not show reduced TE for either the female or the male germline, although it did show a higher pairing frequency. Interestingly, when *ule30/+* mutant is crossed as pollen donor there is a deviation from the 1:1 expected ratio, hemizygous:wild type offspring (p-value: 0.023). Surprisingly, from this cross the offspring that carry the mutation are significantly more abundant than expected. The segregation distortion phenotype persisted for one more generation but unexpectedly the distortion disappeared in the following one. Since ULE30 is located at the promoter region of the *Kan2* gene the distorted segregation we observe could be due to the disruption of the function of *Kan2*. Therefore, we analyzed the segregation of an insertion line (*flag_line*) located between ULE30 and the transcriptional start site of *Kan2* (Figure 4). *flag_line* has a single locus insertion (Supplemental Figure 3). Segregation analysis of the mutant allele resulted in the expected 1:1 ratio (Table 3). Thus, ULE30 seem to act independent of its neighbor gene. In addition, disrupting ULE30 increase the TE of the mutant through the male

gender yet this effect may not last for many generations. Interestingly, no increase in somatic pairing was observed for ULE30.

Table 3. Transmission of the *ule* mutant alleles to the offspring. Deviations from the expected ratios were tested using a chi-square test.

Generation	Female x Male	*ule*/+	+/+	p-value
1st	+/+ x *ule30*/+	431	367	0.023
1st	*ule30*/+ x +/+	433	442	0.761
2nd	+/+ x *ule30*/+	870	770	0.013
3d	+/+ x *ule30*/+	342	324	0.485
	+/+ x *flag_line*/+	328	360	0.222
	flag_line/+ x +/+	427	417	0.73
	+/+ x *ule1*/+	416	376	0.155
	ule1/+ x +/+	512	480	0.309

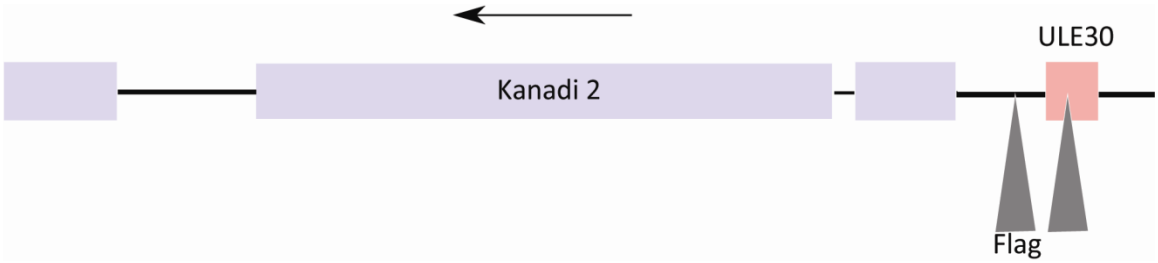


Figure 4. Schematic representation of the genomic region where ULE30 is located

Blue boxes depict coding sequences; arrow direction of transcription, triangles T-DNA insertion lines, red box represents ULE30.

Transmission efficiency of ULE insertion is not distorted in a trisomic background

Since a T-DNA insertion on ULE30 causes altered TE from the expected 1:1 ratio, we argued that ULEs could be involved in addition to somatic pairing to meiotic chromosome pairing. A possible role in meiotic pairing for ULEs could advocate for the chromosome copy counting theory. In meiosis chromosomes do pair and it could be plausible that ULEs have the opportunity to compare their homologs at this stage.

Trisomics have been extensively studied and are characterized by the presence of one extra chromosome in an otherwise diploid background (Blakeslee 1922; Mc Clintock 1929; Rick and Barton 1954; Steinitz-Sears 1963; Koornneef and Van der Veen 1983; Henry et al. 2007). Thus, in *Arabidopsis*, trisomics would have eleven chromosomes instead of ten. Consequently, ULEs of the additional chromosome should in turn have an extra copy which has no homolog to pair with. This irregularity may trigger deleterious events. Therefore, to test whether ULEs are involved in meiotic pairing we decided to follow the transmission efficiency of ULE30. ULE30 was previously shown to affect the TE through the male in a diploid background, thus we wanted to investigate whether the effect is enhanced in a trisomic background, a more sensitized background.

Phenotypes of trisomics is easy to identify based on the additional chromosome (Koornneef and Van der Veen 1983; Henry et al. 2010). Trisomics on chromosome 1 are the easiest to score. They are dwarf plants with dark green, narrow leaves. They are sterile, the stamens are short and no dehiscence occurs (Supplemental Figure 4).

To generate the trisomics, a tetraploid *A. thaliana* (ecotype Col-0) was crossed with a diploid one (Col-0) (Figure 5). The resulting triploid was used in a cross with a diploid that carries the T-DNA insertion on ULE30. From this cross a swarm of aneuploid progeny were produced. Plants that are phenotypically similar to trisomics that have an additional chromosome one and carry the insertion were selected. Since, trisomics on chromosome 1 are male sterile they were used as female plants. Hence, we could not test whether TE is altered through the male as it had been observed before. We used two mutant trisomics on chromosome 1 (Tr1-48, Tr1-53) and used them in crosses with diploid Col-0. Then, the frequency of trisomic genotypes was assessed (Table 4). If insertion on ULE30 cause no effect, the expected ratio of trisomics carrying the insertion relative to trisomics having only wild type alleles would be 2:1. No deviation from the expected ratio was observed when ULE30 copy number is altered in a trisomic background.

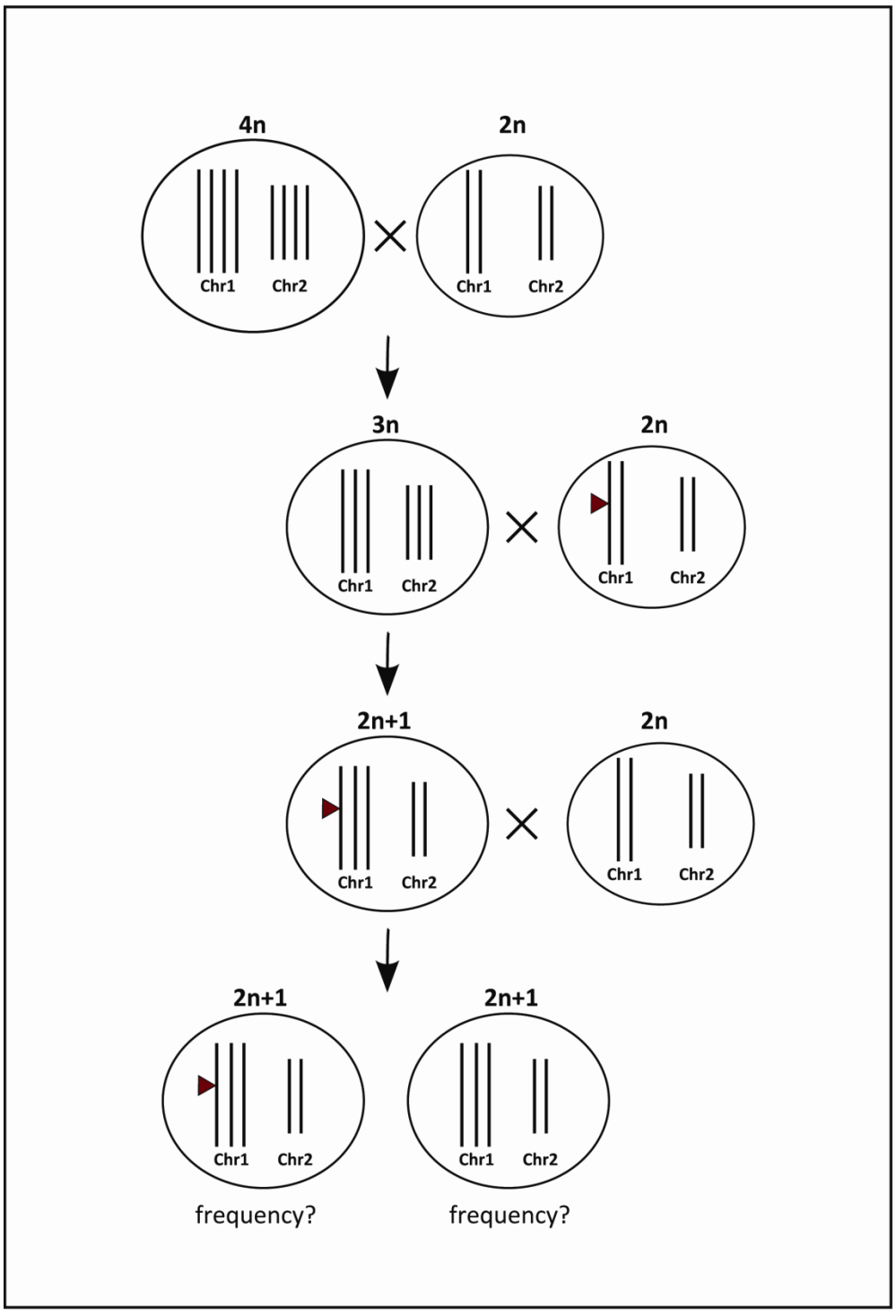


Figure 5. Schematic representation of the genetic crosses performed in order to assess the frequencies of the *A. thaliana* trisomic genotypes

4n: Tetraploid genome; 2n: Diploid genome; 3n: Triploid genome; 2n+1: Trisomic genome. For simplicity purposes only the *Arabidopsis* chromosomes one and two are depicted. Red triangle depicts the T-DNA.

Table 4. Transmission of the *ule30* mutant allele to the offspring in trisomic background. Deviations from the expected ratios were tested using a chi-square test.

+/+/*ule30* x +/+	+/+/*ule30*	+/+/+	Expected ratio (2:1)
Tr1-48 x 2n Col	35	13	0.36
Tr1-53 x 2n Col	18	10	0.68

Similarly, trisomics on chromosome 5 were also produced for the T-DNA insertion on ULE1. However, due to the low frequency of trisomics produced from this cross, segregation analysis could not be performed.

Addition of extra copies of ULEs does not induce chromosomal rearrangements nor affects somatic pairing

As it was reported before, evolutionary forces kept ULEs depleted from segmental duplications, suggesting that deviation from a single copy could have an impact on plant's fitness (Kritsas et al. 2012). Consequently, if ULEs are restricted to be single copy what would be the impact in the plant if we insert more copies of them? In addition, up to now only the effect of single ULE mutants was examined although there might be a certain degree of functional redundancy between the 36 ULEs.

To evaluate this, *Arabidopsis* plants were transformed with a construct carrying four extra copies of ULEs, namely ULE6, ULE7, ULE30 and ULE22 together with some of their flanking regions (13 bp to 20 bp). All of them are located on chromosome one and lie in intergenic regions. ULE6, ULE30, ULE22 are positioned upstream of transcription factors and ULE7 is upstream of a gene whose function is unknown.

Interestingly, five out of thirty independent transformant lines showed semi-sterile phenotype, namely unfertilized ovules due to megaspore mother cell arrest and aborted pollen grains (Figure 6). The female and male meiotic products were aborted ranging from 34% to 50%. That kind of phenotype would be in agreement with ULEs being involved in meiotic pairing. However, all five lines showed no deviation from the Mendelian segregation 3:1 when the offspring of hemizygous plants were screened for antibiotic resistance. For one of these lines, genetic mapping was used to determine the relative position of the insert in the genome. Normally the insert should be linked to one location. In the mutant line, phenotype was genetically linked to mapping markers on two locations in the genome, one in chromosome 2 and chromosome 5 (Materials and Methods). That suggests that the observed phenotype is caused by chromosomal

rearrangement. Likewise, the other four lines showed also Mendelian segregation, thus it was assumed that their phenotype is due to chromosomal translocation as well.

The chromosomal rearrangement phenotype observed could be a byproduct of the T-DNA insertions or it could be triggered by the addition of extra copies of ULEs. To test this, another construct was created with the same T-DNA backbone as before but this time ULE sequences were replaced by part of the GUS gene of equal size. Transformed plants carrying either one or the other construct were grown under the same conditions and were screened for reduced seed set. Although, plants bearing extra copies of ULEs showed a higher frequency of unfertilized ovule phenotype (9/70 transformants) in comparison to control plants (5/89 transformants), the difference was not statistically significant, two-tailed Fisher's exact test, p-value 0.16. This result indicates that addition of extra copies of ULEs does not seem to trigger chromosomal translocation event.

Apart from the transgenic lines that showed chromosomal translocation phenotypes, we identified a single insertion line showing no sterility phenotype. The insertion is located in intergenic space on chromosome four. Segregation analysis of a hemizygous mutant crossed with wild type showed the expected 1:1 ratio. Since, ULEs or the regions that contain them are involved in somatic chromosome pairing; we asked whether the insertion of additional ULEs can induce somatic pairing in the region of insert. A BAC clone spanning the region of insertion was used as a probe in wild type and transgenic plants and the pairing frequency was assessed. Pairing frequency in the nuclei carrying extra ULE copies (15/110) was not statistically different from the wild type region (6/80) (Table 5).

Genetic and cytogenetic analysis of plants bearing extra copies of ULEs prompts that increasing the number of ULEs does not appear to affect the plant's development.

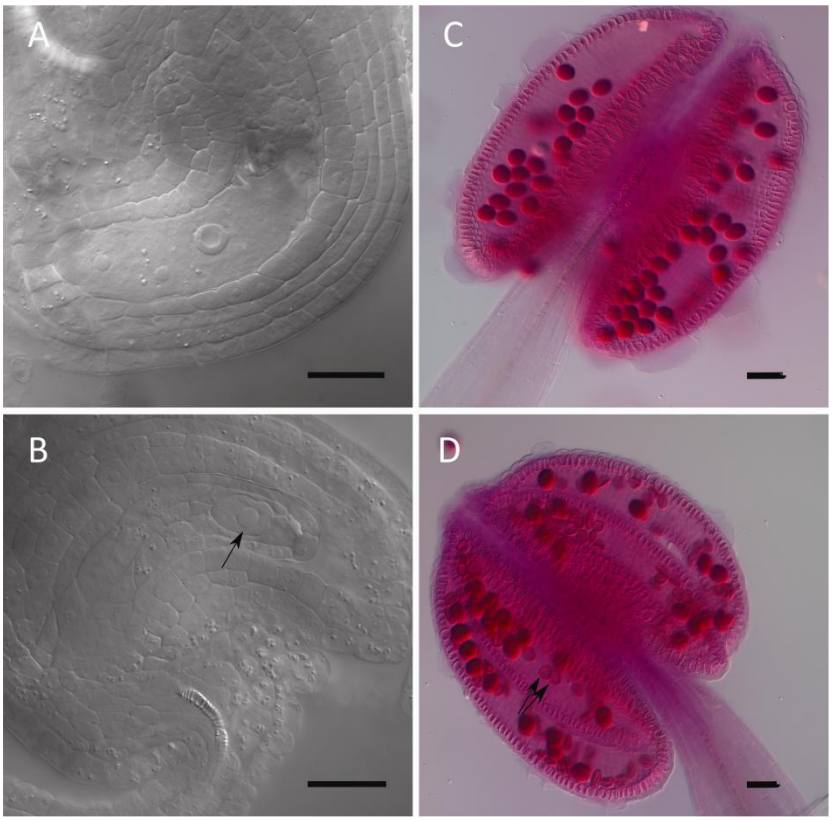


Figure 6. Aborted meiotic products of trangenic *A. thaliana* plants carrying extra copies of ULEs

A. Wild-type mature embryo sac. B. Transgenic embryo sac where megaspore mother cell is aborted. C. Wild-type mature pollen grains. D. Transgenic aborted pollen grains. Black arrows indicate the aborted meiotic products. Bar = 200 μm.

Table 5. Pairing frequency of a homologous chromosome region acquired extra copies of ULEs in comparison to the pairing frequency of the same region without additional copies. Differences between the two pairing frequencies were tested with two-tailed Fisher's exact test.

BAC	Chromosome	Nuclei	Pairing frequency (%)	Two-tailed p-value
ExULEs T28I19	4	110	13.63	0.24
Wild type T28I19	4	80	7.5	

ExULEs is the the region gaining additional ULEs, T28I19 is the BAC clone which was used as a probe in the FISH assay.

Artificial *de novo* silencing of ULEs does not affect the fitness of the plant

According to genome wide DNA methylation data from 5-wk-old plants and flower buds, most of the ULEs do not appear to be methylated in any sequence context (CG, CHG, CHH) and those who are, are not extensively methylated (Kritsas et al. 2012). We reasoned that absence of DNA methylation at ULE loci could have a functional implication; therefore we aimed at altering the DNA methylation status of four ULEs (ULE6, ULE7, ULE30, and ULE22) with a transgene that would trigger the RNA-directed DNA methylation pathway (RdDM).

Plants have evolved a unique way to induce transcriptional gene silencing. In the RdDM pathway, 24-nt small RNAs (siRNAs) target the DNA methylation machinery in specific loci which display siRNA-DNA homology (Wassenegger et al. 1994; Matzke et al. 2009; Law and Jacobsen 2010). RdDM induces methylation in all sequence contexts.

In our study, we followed an RNA-dependent DNA methylation strategy, by using the pHellsgate12 transgene (Helliwell 2003; Kinoshita et al. 2007). Hellsgate vectors are Gateway compatible and the insertion of the four ULE sequences is achieved in a single recombination step. ULEs were inserted in both forward and inverse orientation separated by an intron (Supplementary Figure 6). ULEs expression is driven by the strong constitutive promoter 35S and subsequently the corresponding hairpin RNA should be produced and target the methylation of the four ULEs.

Transformed plants were then tested to check if *de novo* methylation did occur on the four ULE regions by using the McrBC-PCR assay. McrBC is a restriction enzyme that cleaves between methylated cytosine residues but not unmethylated DNA (Sutherland et al. 1992). Therefore, the region which is methylated will not be amplified by PCR by primers that flank the region of interest. We used this assay to test whether ULE6, ULE30, ULE7 and ULE22 are methylated under the control of the Hellsgate12 transgene (Figure 7). In this assay, the methylated copia-like transposon *TA2* was used as a positive control (Vaughn et al. 2007). Failure to amplify digested DNA in the transgenic samples ULE6 and ULE22 indicates that these regions became methylated. For ULE30 and ULE7 methylation status is ambiguous and it seems there is no difference between the treated transgenic DNA with the controls, suggesting that these two regions were already methylated in the wild type background. Hence, two out of four ULEs were methylated.

Next, transgenic lines were tested for sterility phenotypes, unfertilized ovules and pollen viability. No statistical difference was observed in comparison to the wild type. Our

findings suggest that although the methylation status of at least two ULEs was altered, yet there was no obvious impact on the fitness of the plant.

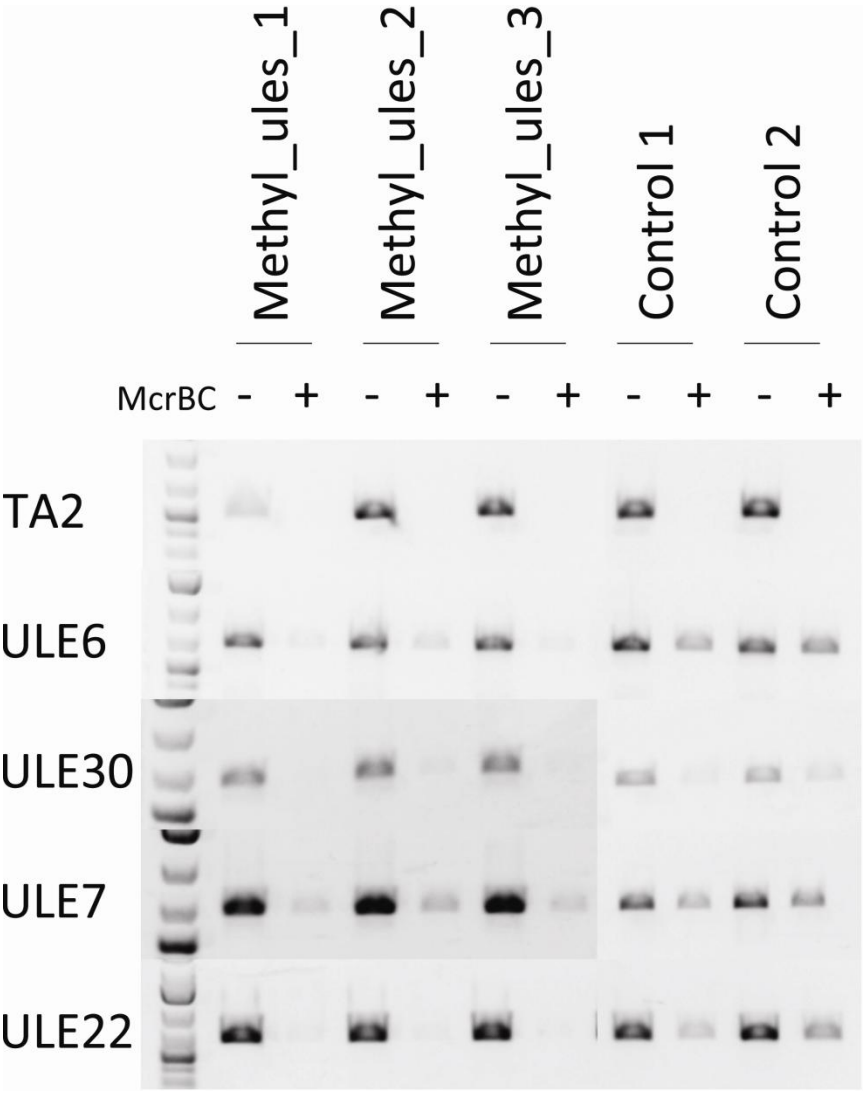


Figure 7. *De novo* methylation of ULEs

PCR amplification of undigested DNA McrBC (-) and digested DNA McrBC (+). Failure to amplify DNA in the McrBc (+) samples suggests that these regions are methylated. Amplification of transposon *TA2* was used as a positive control for methylated DNA. Methyl_ULEs_1,2,3 are samples from RdDM transgenes, control 1,2 are samples from wild type *A. thaliana* Col-0.

Material and Methods

Preparation of nuclei, probe labeling and FISH

Three to four young rosette leaves of *A. thaliana* accession Col-0 were fixed for 20 min in 4% formaldehyde (Sigma F-1635) in TRIS buffer (10 mM TRIS-HCl, 10 mM Na_2EDTA, 100 mM NaCl, pH 7,5, 0.1% Triton X-100). Fixative buffer was rinsed two times with TRIS buffer. Leaves were then homogenized in nuclei isolation buffer (15 mM TRIS-HCl, 2 mM Na_2EDTA, 0.5 mM spermin, 80 mM KCl, 20 mM NaCl, 15 mM 2-mercaptoethanol, 0.1% Triton X-100). Suspended nuclei were then passed through 30 µm mesh filter and stained with 1 µl Sytox Blue Dead Cell stain (Invitrogen). Nuclei were sorted according to their ploidy with a flow sorter, BD FACSAria IIIu BL1 sorter. Approximately, 1'000 diploid nuclei were sorted on microscopic slides in a drop of "sucrose pillow" (100 mM TRIS-HCl, 50 mM KCl, 2 mM $MgCl_2$, 0.05% Tween-20, 5% sucrose, filter sterilize), air-dried and stored in -20^{0}C until use.

Bacterial artificial clones (BACs) were ordered from the Arabidopsis Biological Resource Center, (abrc.osu.edu). Single copy clones were selected according to the DNA-DNA dot blot hybridizations of *A. thaliana* genomic DNA (Lysak et al. 2003). All BACs were verified by PCR. BAC clones used are listed in Supplementary Table 1. Alkaline lysis method was used to isolate DNA from BACs. For probe labeling, the nick translation method was followed: 900 ng of DNA that contains ULE sequences was labeled with the Dig-nick translation mix (Roche) and the neighboring DNA region with Biotin-nick translation mix (Roche). When the labeled fragments were in the range of 200-500 nucleotides the reaction was stopped, by adding 1 µl 0.5 M EDTA (pH 8.0) and heating to 65^{0}C for 10 min. Probes were then cleaned with QIAquick nucleotide removal kit (Qiagen). Dig and biotin labeled probes were mixed together and ddH_2O was added up to 50 µl, followed by ethanol precipitation. The air-dried probe was diluted in 10 µl HB50, heated at 42^{0}C (15 min) and 10 µl of 20% dextran sulfate in HB50 was added. Probe was then denatured at 75^{0}C (15 min) and rest on ice.

Prior to hybridization, slides were rinsed with 2x SSC (5 min), fixed in 1% formaldehyde in PBS (5 min), rinsed in 1x PBS (5 min), treated slides with pepsin (Sigma F1887) for 90 sec, post-fixed nuclei in 1% formaldehyde in 1x PBS (10 min), rinsed in 1x PBS (2x5 min) and dehydrated in 70, 90 and 100% ethanol (2 min each) and air-dried. Slides then were treated with RNase (100 µl of 100 µg/mL RNase A in 2X SSC) for 30 min at 37^{0}C, rinsed with 2x SSC (2x5 min), 1x PBS (5 min), dehydrated 70, 90 and 100% ethanol (2

min each) and air-dried.

For hybridization, 20 µl of probe was added to each slide. Probe and chromosomal DNA were denatured at 80^0C (2 min) on a heating block. Slides were put in a humid chamber and incubated for at least 18 hours at 37^0C.

Post-hybridization washes were performed at 42^0C. Slides were washed with SF50 (3x5 min), followed by 2xSSC (2x5 min), 4T (4xSSC, 0.05% Tween-20) (5 min). For the detection, all washes were performed under stringent conditions at 42^0C. Biotin labeled probes were detected with Texas Red Avidin DCS (Vector Labs, A-2016; 1:1'000) and biotinylated Anti-Avidin D (Vector Labs, BA-0300; 1:250). Dig labeled probes were detected with mouse anti-digoxigenin (Roche; 1:250) and goat anti-mouse conjugated with Alexa-488 (Life Technologies; 2.5:1'000). After the post-hybridization washes, 100 µl of blocking solution was added upon the slides (Vector Labs, MB-1220; 30 min at 37^0C), followed by 4T (2x5 min). Texas Red diluted in blocking solution was added (30 min at 37^0C) and rinsed by 4T (2x5 min) and TNT (5 min). Anti-Avidin and mouse anti-digoxigenin antibodies in TNB were added (30 min at 37^0C), followed by TNT (3x5 min). Texas-Red and goat anti-mouse~Alexa-488 in TNB was added (30 min at 37^0C), rinsed in TNT (3x5 min). Slides were then dehydrated in 70, 90 and 100% ethanol (2 min each) and air-dried and a small drop of Vectashield (Vector Labs, H-1200) was mounted. Slides were stored at 4^0C until microscopy analysis.

The fluorescent signals from the FISH-treated nuclei were visualized under an epifluorescence microscope (DM6000 Leica) equipped with filters for detection of DAPI (exitation: 340-380 nm), Alexa-488 (excitation: 480/40 nm), and Texas Red fluorescence (excitation: 560/40 nm), and pictures were taken with a Leica DFC350FXR2 digital camera and analyzed with Leica application suite software.

Plant material, growth conditions, mutant genotyping and trisomics

Arabidopsis thaliana plants were grown in growth chambers in plastic pots filled with ready to use soil (Einheitserde). After sowing, plants were kept at 4^0C for two days. Growing conditions were $21\text{-}23^0$C, 65% humidity, with a 16hr light / 8hr dark photoperiod regime at ~75 µmol m^{-2} s^{-1}.

ule6 (SAIL_1265_F06; N879048) and *ule30* (SAIL_896_G06; N877830) mutant allele seeds were obtained from The European Arabidopsis Stock Centre (arabidopsis.info). *ule1* (GABI_862D05) seeds were obtained from GABI-Kat (www.gabi-kat.de).

Genotyping assays for *ule6, ule30, ule1* alleles were performed with the following primer pairs: *ule6* (5'-TCCCAAAGTCTCACCACTCAC-3') and (5'-GCCTTTTCAGAAATGGATAAATAGCCTTGCTTCC-3'), *ule30* (5'-GCTTAACCTCTTACGGCCATC-3') and (5'-GCCTTTTCAGAAATGGATAAATAGCCTTGCTTCC-3'), *ule1* (5'-CGAATGACTGTTAAGGCTTCG-3') and (5'-ATATTGACCATCATACTCATTGC-3').

Manual crosses were performed as previously described (Boisson-Dernier et al. 2008); closed flower buds from late stage 12 were emasculated and then manually pollinated 48hr later from pollen donor flowers under a dissecting microscope.

Trisomic plants were produced from tetraploid *A. thaliana* Col-0 plants as described in Figure 5. Tetraploid seeds were kindly provided from Luca Comai lab. Genome content of tetraploid and triploid plants was verified with flow cytometry (Beckman-Coulter, Quanta SC MPL) using diploid Col-0 plants as control of known genome content. Identification of aneuploid individuals was based on the phenotypic characteristics described before (Koncz et al. 1992; Henry et al. 2010; Isabelle M. Henry personal communication) and always in comparison with diploid plants grown at the same time under identical conditions.

Addition of extra copies of ULEs and *de novo* methylation

Binary vector bearing extra copies of ULEs (Supplementary Figure 5) or the partial sequence of GUS gene were introduced into *Agrobacterium tumefaciens* strain GV3101 by electroporation which was subsequently used to transform wild type Col-0 plants by floral dipping. ULE sequences together with their flanks inserted are provided in Supplementary Table 2.

Rough mapping of the chromosomal rearrangement was performed on *kk248-6/+* x Ler F2 population using accession specific polymorphisms. KK248-6 line showed aborted meiotic products after the insertion of four ULEs. Semi-sterile plants (n=40) were separated for mapping the *kk248-6* allele. Two genetic markers mapped on chromosome two, CER458319 and CER4602 (Salathia et al. 2007) and three from chromosome five, CIW9, CER454487, CER457837 (Berendzen et al. 2005; Salathia et al. 2007) are segregating with the semi-sterile phenotype.

DNA methylation of the ULEs was assessed by PCR amplification of DNA that has been treated with the methylation specific restriction enzyme McrBC. Previously, genomic

DNA was extracted with DNeasy Plant kit (Qiagen). 100 ng of genomic DNA were treated with 10 U of McrBC (New England Biolabs) overnight at 37^0C, followed by ethanol precipitation. Control samples were treated the same without the enzyme. PCR amplification was performed on 10 ng/μl digested DNA for 28 cycles, at 60^0C annealing temperature. Primer sequences for the *TA2* positive control (Lodha et al. 2013) and ULEs are indicated on Supplementary Table 3.

For the aborted meiotic products phenotype, ovule clearings were performed with Hoyer's solution (Lewis 1954) and pollen viability with Alexander's staining (Alexander 1969). The Leica DMR microscope was employed for microscopy analysis.

Discussion

NcUCEs/ULEs are extremely conserved non-coding sequences whose purpose of existence remains a riddle. In this study, we are exploring the mechanism underlying the function of the ULEs, the plant counterparts of conserved non-coding sequences. Our hypothesis is that ULEs are part of a homolog pairing mechanism and act as agents of genome integrity by making sure that the correct copy of chromosomes is conserved. The chromosome copy counting model (Derti et al. 2006) is supported by the genome wide distribution and uniqueness of the ULEs as well as by the fact that ULEs together with ncUCEs are absent from duplicated regions (Chiang et al. 2008; Kritsas et al. 2012). We tried to explore the counting hypothesis by employing cytogenetic and genetic tools.

Homologous pairing in ULE regions occurs more often than random

We argued that if ULEs are part of a chromosome copy counting mechanism this should be mediated through homologous DNA interactions. Therefore, we investigated whether homologous chromosomes at the ULE regions show a higher pairing frequency. Our FISH assay on leaf somatic interphase nuclei revealed that in seven out of nine ULE regions pairing frequency was higher when compared with other regions lying on the same chromosome. In eukaryotic somatic nuclei, chromosomes occupy distinct space, called chromosome territories (CTs) and normally they don't intermingle with each other (Cremer et al. 2001). *Arabidopsis* species have similar organization and in fact homologous chromosome pairing seem to be random (Pecinka et al. 2004; Berr et al. 2006). Nevertheless, there are exceptions to this trend. It is shown, that nucleolus organizing regions (NORs) associate more often than expected because of their attachment to the single nucleolus (Pecinka et al. 2004; Berr et al. 2006). Increased homologous chromosome pairing has been also observed in the case of regions that have long stretches of tandem repeats (Pecinka et al. 2005; Watanabe et al. 2005; Jovtchev et al. 2011). Our results indicate that ULEs belong to regions that are also part of the exemption and exhibit increased pairing.

Except for somatic are ULEs involved in a copy counting mechanism during meiosis? The best way to resolve this is to study ULE pairing frequency at zygotene stage of meiosis I. At this stage homologous chromosomes do actually pair in order to ensure that they segregate faithfully to the germ cells (Ronceret and Pawlowski 2010). It would be

intriguing to see whether ULEs are part of pairing initiation sites just before synapsis. Interestingly, in wheat, homologous chromosomes associate non-randomly pre-meiotically in meiocytes (Aragón-Alcaide et al. 1997). However, quantifying FISH signal at this stage is technically challenging. Therefore, we followed a genetic approach to address whether ULEs are involved in meiotic pairing.

Are ULEs involved in meiotic pairing? Are they indispensable?

In mice, deletion of four ncUCEs failed to reveal any critical abnormalities (Ahituv et al. 2007). In *Arabidopsis* homozygous insertional mutant on ULE6 yielded phenotypes similar to the *monopteros* gene, which is located 700 bp downstream of ULE6. This prompts, that like the animal ncUCEs, some of the ULEs may act as enhancers. The putative enhancer-like activity of the ULEs is not in conflict with the copy counting hypothesis, since both functions require pairing-mediated mechanisms.

Counting hypothesis suggests that absence of both copies of ncUCEs/ULEs (homozygosity) can be less detrimental since there is no opportunity for sequence comparison as oppose to a heterozygous status (Derti et al. 2006). Interestingly, the consequence of hemizygous *ule30/+*, thus perturbation of just one copy of ULE30, is that the mutant allele is transmitted to the offspring in excess of the expected Mendelian proportion of 50%. It seems that although ULEs are under negative selection, perturbation of one ULE gives a selective advantage. One explanation could be that in a hemizygous background the unpaired, wild-type allele of ULE30 triggers a deleterious signal, hence less copies of the wild-type allele are transmitted to the next generation.

The segregation distortion phenotype of *ule30* mutant allele is lost after a few generations. Bearing the copy counting mechanism in mind, this observation suggests that when large alterations of the sequence of ULEs is taking place, very fast the homologous ULE comparison process is compromised. Thus, heavily mutated sequences are eventually excluded from counting activities. In our lab conditions phenotype disappears after just two generations but since T-DNA seeds are derived from the European stock center, it is possible that they were propagated for a few generations there as well before arriving in our hands. Hence, it is not known how many generations are required to render ULEs ineffective.

In view of the skewed frequency of the *ule30* mutant, we took advantage of the ability of plants to tolerate aneuploidy and asked whether the extent of distortion is enhanced in a

trisomic situation. Trisomics correspond to a more sensitized background. Although the lack of distortion in trisomics advocates against the meiotic pairing function of ULE30, it could be possible that the direction of crossing *ule30* is important. In diploid plants the TE phenotype occurs when the *ule30* allele is inherited from the male parent. Trisomics on chromosome one are male sterile plants subsequently *ule30* allele was used in crosses as a female. It could be conceivable that the mechanism of comparison may be gender specific and this process can be observed only when mutated *ule30* is transmitted through pollen.

Chromosome copy counting model via sequence comparison predicts that additional copies of ULEs can have similar deleterious effect as their disruption. We argued that the chromosomal rearrangement phenotypes we observed could be due to the increased proximity of chromosomal loci that have the extra ULEs (Nikiforova et al. 2000; Roix et al. 2003; Cavalli 2007; Lin et al. 2009). However, there was no statistical difference with the control plants. Furthermore, the pairing frequency of a locus on chromosome four bearing extra ULEs remained unchanged. However, since all ULEs inserted belong to chromosome one, the effect could be chromosome specific. It would be interesting to investigate whether the pairing frequency of loci belonging to chromosome one is changing when they acquire more ULE copies.

Pairing and comparison of the ULEs could be mediated by DNA binding proteins. Since ULEs are not methylated, we asked whether this protein-DNA interaction is compromised when the methylation status of ULEs is altered. In nature there is a number of transcription factors which are particularly sensitive to DNA methylation (Tate and Bird 1993). We were successful in inducing the methylation of at least two ULEs. However, no obvious developmental defects were observed, suggesting that ULEs function is immune to DNA methylation.

The robust nature of ULEs (high conservation, purifying selection, single copy), implies that are essential elements for plants. Although, an insertion on ULE30 causes segregation distortion, the effect of altering the copy number of these elements is not deleterious. It may be that disturbing the function of ULEs has a long-term effect and its consequence can be traced on longer evolutionary time. Further, it may be that ULEs act in a redundant and/or cell specific manner and their function was difficult to decipher with our experimental settings.

Acknowledgements

We thank Dr. Christof Eichenberger for introduction to DM6000 Leica microscopy. Dr. Afif Hedhly for the ovule clearing and Alexander staining protocol. Dr. Frédérique Pasquer for the ploidy analysis of *A. thaliana* Col-0 polyploid nuclei. Dr. Aurelien Boisson-Dernier for critically reading the manuscript. Dr. Isabelle Henry (UC Davis) for useful discussions about the phenotype of trisomics. Dr. C-ting Wu (Harvard Medical School) for thought provoking discussions about the nature of the ULEs.

References

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio L a, Rubin EM. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol* **5**: e234.

Alexander M. 1969. Differential staining of aborted and nonaborted pollen. *Stain Technol* **44**: 117–122.

Aragón-Alcaide L, Reader S, Beven A, Shaw P, Miller T, Moore G. 1997. Association of homologous chromosomes during floral development. *Curr Biol* **7**: 905–8.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* **304**: 1321–5.

Berendzen K, Searle I, Ravenscroft D, Koncz C, Batschauer A, Coupland G, Somssich IE, Ulker B. 2005. A rapid and versatile combined DNA/RNA extraction protocol and its application to the analysis of a novel DNA marker set polymorphic between Arabidopsis thaliana ecotypes Col-0 and Landsberg erecta. *Plant Methods* **1**: 4.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**: 315–26.

Berr A, Pecinka A, Meister A, Kreth G, Fuchs J, Blattner FR, Lysak M a, Schubert I. 2006. Chromosome arrangement and nuclear architecture but not centromeric sequences are conserved between Arabidopsis thaliana and Arabidopsis lyrata. *Plant J* **48**: 771–83.

Blakeslee AF. 1922. Variations in Datura due to changes in chromosome number. *Am Soc Nat* **56**: 16–31.

Boisson-Dernier A, Frietsch S, Kim T-H, Dizon MB, Schroeder JI. 2008. The peroxin loss-of-function mutation abstinence by mutual consent disrupts male-female gametophyte recognition. *Curr Biol* **18**: 63–8.

Calin G a, Liu C, Ferracin M, Hyslop T, Spizzo R, Sevignani C, Fabbri M, Cimmino A, Lee EJ, Wojcik SE, et al. 2007. Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* **12**: 215–29.

Cavalli G. 2007. Chromosome kissing. *Curr Opin Genet Dev* **17**: 443–50.

Chiang CWK, Derti A, Schwartz D, Chou MF, Hirschhorn JN, Wu C-T. 2008. Ultraconserved elements: analyses of dosage sensitivity, motifs and boundaries. *Genetics* **180**: 2277–93.

Cremer M, von Hase J, Volm T, Brero A, Kreth G, Walter J, Fischer C, Solovei I, Cremer C, Cremer T. 2001. Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. *Chromosom Res* **9**: 541–67.

Curtis MD, Grossniklaus U. 2003. A Gateway Cloning Vector Set for High-Throughput Functional Analysis of Genes in Planta. *Plant Physiol* **133**: 462–469.

Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV, et al. 2002. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420**: 578–82.

Derti A, Roth FP, Church GM, Wu C. 2006. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet* **38**: 1216–20.

Feng J, Bi C, Clark BS, Mady R, Shah P, Kohtz JD. 2006. The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev* **20**: 1470–84.

Ghanem N, Jarinova O, Amores A, Long Q, Hatch G, Park BK, Rubenstein JLR, Ekker M. 2003. Regulatory Roles of Conserved Intergenic Domains in Vertebrate Dlx Bigene Clusters. *Genome Res* **13**: 533–543.

Glazov E a, Pheasant M, McGraw E a, Bejerano G, Mattick JS. 2005. Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res* **15**: 800–8.

Hardtke CS, Berleth T. 1998. The Arabidopsis gene MONOPTEROS encodes a transcription factor mediating embryo axis formation and vascular development. *EMBO J* **17**: 1405–11.

Helliwell C. 2003. Constructs and methods for high-throughput gene silencing in plants. *Methods* **30**: 289–295.

Henry IM, Dilkes BP, Comai L. 2007. Genetic basis for dosage sensitivity in Arabidopsis thaliana. *PLoS Genet* **3**: e70.

Henry IM, Dilkes BP, Miller ES, Burkart-Waco D, Comai L. 2010. Phenotypic consequences of aneuploidy in Arabidopsis thaliana. *Genetics* **186**: 1231–45.

Jovtchev G, Borisova BE, Kuhlmann M, Fuchs J, Watanabe K, Schubert I, Mette MF. 2011. Pairing of lacO tandem repeats in Arabidopsis thaliana nuclei requires the presence of hypermethylated, large arrays at two chromosomal positions, but does not depend on H3-lysine-9-dimethylation. *Chromosoma* **120**: 609–19.

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D. 2007. Human genome ultraconserved elements are ultraselected. *Science* **317**: 915.

Kinoshita Y, Saze H, Kinoshita T, Miura A, Soppe WJJ, Koornneef M, Kakutani T. 2007. Control of FWA gene silencing in Arabidopsis thaliana by SINE-related direct repeats. *Plant J* **49**: 38–45.

Koornneef M, Van der Veen JH. 1983. Trisomics in Arabidopsis thaliana and the location of linkage groups. *Genetica* **61**: 41–46.

Kritsas K, Wuest SE, Hupalo D, Kern AD, Wicker T, Grossniklaus U. 2012. Computational analysis and characterization of UCE-like elements (ULEs) in plant genomes. *Genome Res* 2455–2466.

Law J a, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* **11**: 204–20.

Lee TI, Jenner RG, Boyer L a, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K, et al. 2006. Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**: 301–13.

Lewis AA. 1954. Hoyer ' s Solution as a Rapid Permanent Mounting Medium for Bryophytes. *Bryologist* **57**: 242–244.

Licastro D, Gennarino V a, Petrera F, Sanges R, Banfi S, Stupka E. 2010. Promiscuity of enhancer, coding and non-coding transcription functions in ultraconserved elements. *BMC Genomics* **11**: 151.

Lin C, Yang L, Tanasa B, Hutt K, Ju B, Ohgi K, Zhang J, Rose DW, Fu X-D, Glass CK, et al. 2009. Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. *Cell* **139**: 1069–83.

Lodha M, Marco CF, Timmermans MCP. 2013. The ASYMMETRIC LEAVES complex maintains repression of KNOX homeobox genes via direct recruitment of Polycomb-repressive complex2. *Genes Dev* **27**: 596-601.

Lysak M a, Pecinka A, Schubert I. 2003. Recent progress in chromosome painting of Arabidopsis and related species. *Chromosom Res* **11**: 195–204.

Matzke M, Kanno T, Daxinger L, Huettel B, Matzke AJM. 2009. RNA-mediated chromatin-based silencing in plants. *Curr Opin Cell Biol* **21**: 367–76.

Mc Clintock B. 1929. A Cytological and Genetical Study of Triploid Maize. *Genetics* **14**: 180–222.

Nikiforova MN, Stringer JR, Blough R, Medvedovic M, Fagin JA, Nikiforov YE. 2000. Proximity of Chromosomal Loci That Participate in Radiation-Induced Rearrangements in Human Cells. *Science* **290**: 138–141.

Ovcharenko I. 2008. Widespread ultraconservation divergence in primates. *Mol Biol Evol* **25**: 1668–76.

Pecinka A, Kato N, Meister A, Probst A V, Schubert I, Lam E. 2005. Tandem repetitive transgenes and fluorescent chromatin tags alter local interphase chromosome arrangement in Arabidopsis thaliana. *J Cell Sci* **118**: 3751–8.

Pecinka A, Schubert V, Meister A, Kreth G, Klatte M, Lysak M a, Fuchs J, Schubert I. 2004. Chromosome territory arrangement and homologous pairing in nuclei of Arabidopsis thaliana are predominantly random except for NOR-bearing chromosomes. *Chromosoma* **113**: 258–69.

Pennacchio L a, Ahituv N, Moses AM, Prabhakar S, Nobrega M a, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**: 499–502.

Poulin F, Nobrega M a, Plajzer-Frick I, Holt A, Afzal V, Rubin EM, Pennacchio L a. 2005. In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics* **85**: 774–81.

Rick CM, Barton DW. 1954. Cytological and genetical identification of the primary trisomics of the tomato. *Genetics* **39**: 640–666.

Roix JJ, McQueen PG, Munson PJ, Parada L a, Misteli T. 2003. Spatial proximity of translocation-prone gene loci in human lymphomas. *Nat Genet* **34**: 287–91.

Ronceret a, Pawlowski WP. 2010. Chromosome dynamics in meiotic prophase I in plants. *Cytogenet Genome Res* **129**: 173–83.

Salathia N, Lee HN, Sangster T a, Morneau K, Landry CR, Schellenberg K, Behere AS, Gunderson KL, Cavalieri D, Jander G, et al. 2007. Indel arrays: an affordable alternative for genotyping. *Plant J* **51**: 727–37.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–50.

Steinitz-Sears LM. 1963. Chromosome studies in Arabidopsis thaliana. *Genetics* **48**: 483–490.

Stephen S, Pheasant M, Makunin I V, Mattick JS. 2008. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol Biol Evol* **25**: 402–8.

Stormo GD. 2000. DNA binding sites: representation and discovery. *Bioinformatics* **16**: 16–23.

Sutherland E, Coe L, Raleigh EA. 1992. McrBC : a Multisubunit Restriction Endonuclease. *J Mol Biol* **225**: 327–348.

Tate PH, Bird a P. 1993. Effects of DNA methylation on DNA-binding proteins and gene expression. *Curr Opin Genet Dev* **3**: 226–31.

Vaughn MW, Tanurdzić M, Lippman Z, Jiang H, Carrasquillo R, Rabinowicz PD, Dedhia N, McCombie WR, Agier N, Bulski A, et al. 2007. Epigenetic natural variation in Arabidopsis thaliana. *PLoS Biol* **5**: e174.

Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G. 2007. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biol* **8**: R15.

Visel A, Prabhakar S, Akiyama J a, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio L a. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40**: 158–60.

Viturawong T, Meissner F, Butter F, Mann M. 2013. A DNA-Centric Protein Interaction Map of Ultraconserved Elements Reveals Contribution of Transcription Factor Binding Hubs to Conservation. *Cell Rep* 1–15.

Wang J, Lee AP, Kodzius R, Brenner S, Venkatesh B. 2009. Large number of ultraconserved elements were already present in the jawed vertebrate ancestor. *Mol Biol Evol* **26**: 487–90.

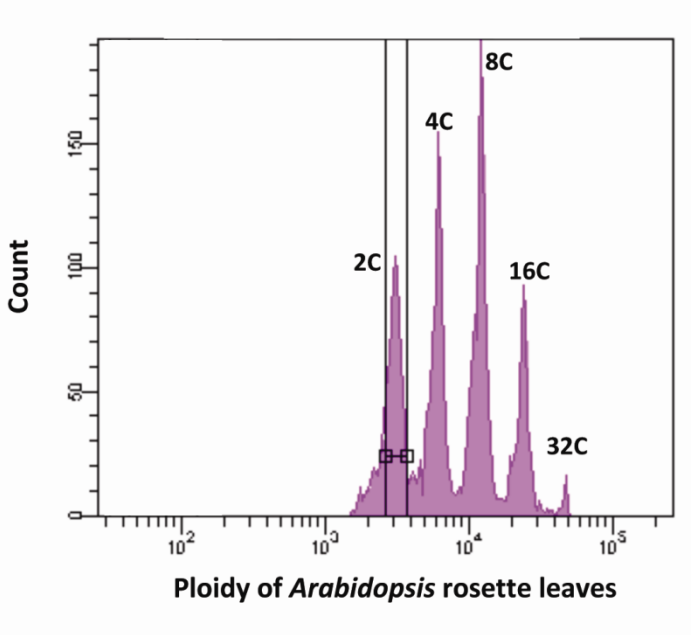
Wassenegger M, Heimes S, Riedel L, Sänger HL. 1994. RNA-directed de novo methylation of genomic sequences in plants. *Cell* **76**: 567–76.

Watanabe K, Pecinka A, Meister A, Schubert I, Lam E. 2005. DNA hypomethylation reduces homologous pairing of inserted tandem repeat arrays in somatic nuclei of Arabidopsis thaliana. *Plant J* **44**: 531–40.

Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**: e7.

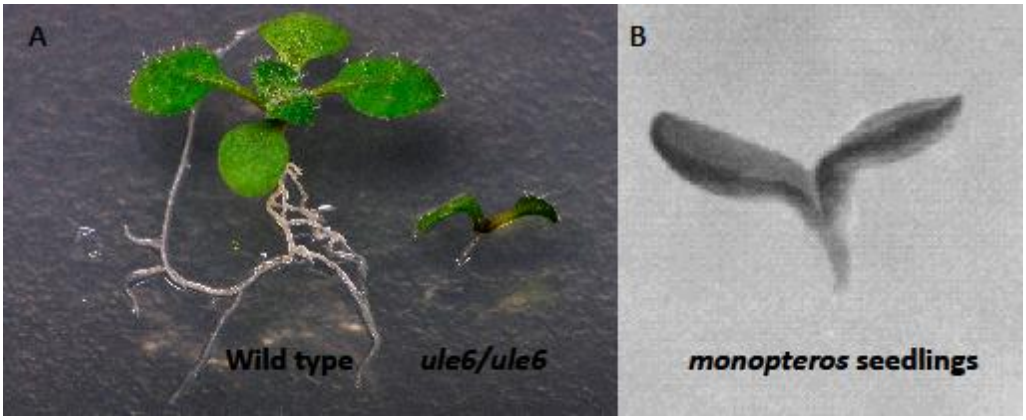
Supplementary materials

Supplementary Figures



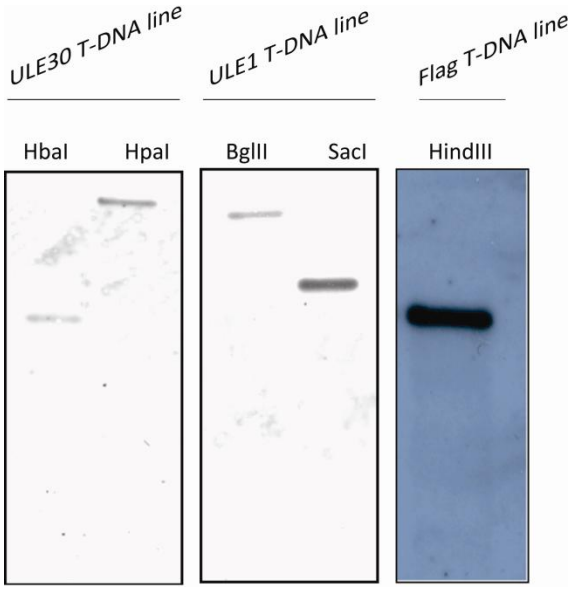
Supplemental Figure 1. **Flow sorting of 2C *Arabidopsis* rosette leaf nuclei**

Nuclei were stained with Sytox Blue Dead Cell Stain (Invitrogen) and the ploidy of the nuclei suspension was determined according to their ploidy. The two vertical lines indicate the fraction of the diploid (2C) nuclei that was collected.



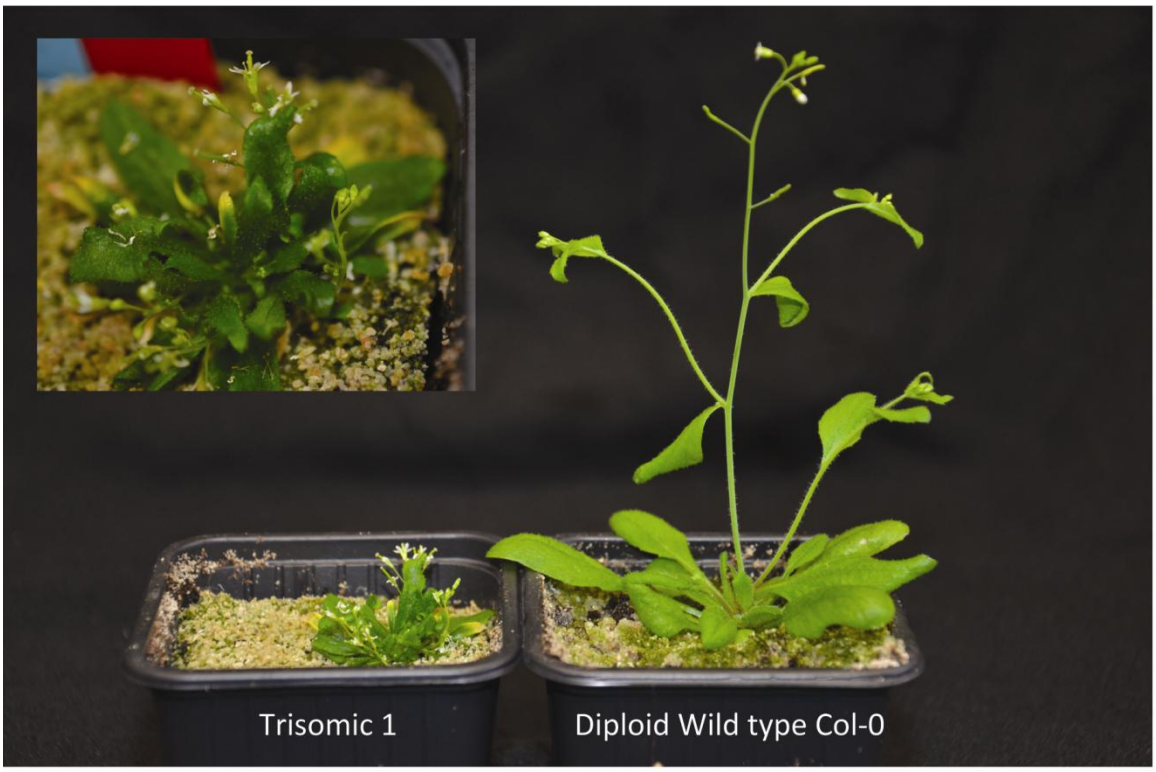
Supplemental Figure 2. **Phenotype of ULE6 T-DNA insertional mutant**

(A) Homozygous *ule6* seedling phenotype. Basal part of the seedling (hypocotyl, root) is dramatically reduced. Phenotype is similar to *monopteros* mutant. (B) *monopteros* mutant seedling phenotype. The basal part of the seedling is eliminated. *monopteros* seedling image is adapted from (Berleth and Jürgens 1993).



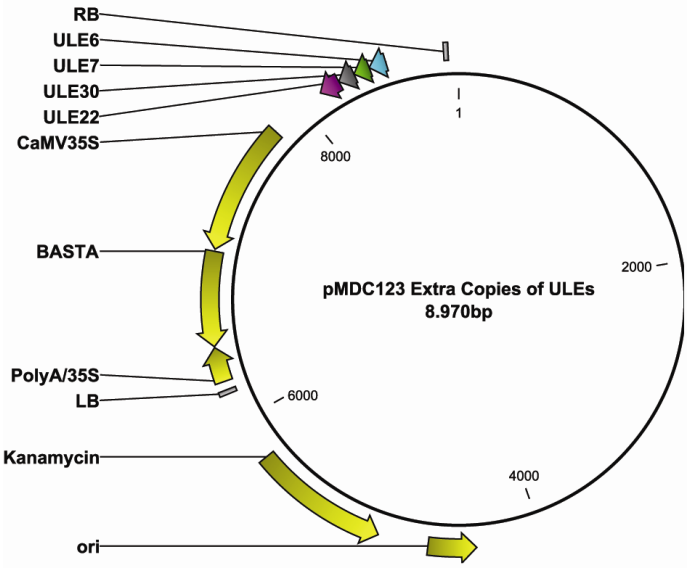
Supplemental Figure 3. **Southern blot hybridizations on ULE T-DNA insertion lines**

Genomic DNA samples were prepared from ULE30, ULE1, Flag T-DNA lines. Genomic DNA was digested for (A) ULE30 with HbaI and HpaI, (B) ULE1 with BglII and SacI, (C) Flag line with HindIII. All samples were size fractionated by gel electrophoresis. After transfer to a nylon membrane, DNA samples were hybridized with probes corresponding to 250 bp of the BASTA resistance gene for ULE30 and Flag line and a probe corresponding to 356 bp of the sulfadiazine resistance gene for the ULE1 T-DNA line.



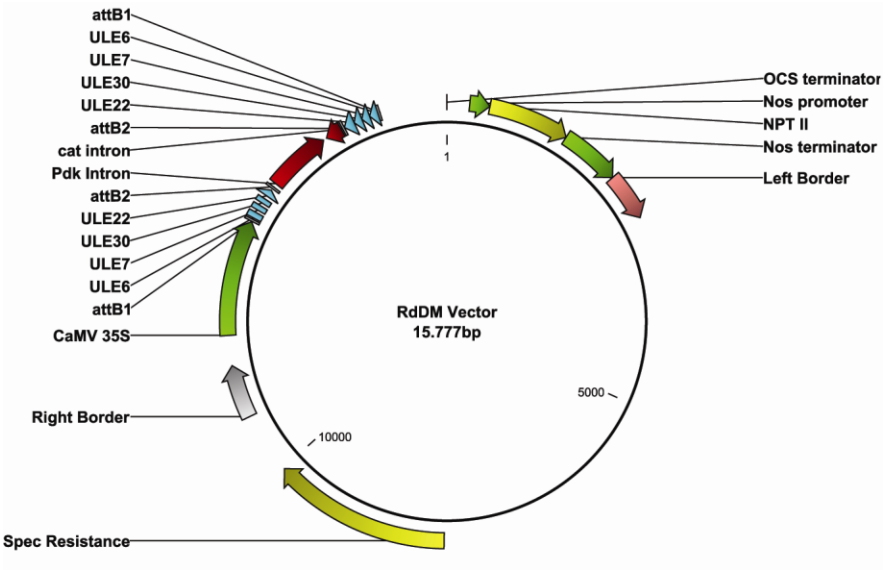
Supplemental Figure 4. **Phenotype of *A .thaliana* plants carrying an extra chromosome 1 (Trisomic 1)**

Trisomics on chromosome one are dwarf plants. They have dark green, narrow leaves. They are sterile plants. Flower organs are thin, the pistil is protruding and the stamens are short and don't dehisce.



Supplemental Figure 5. Map of the vector used to add extra copies of ULEs

The Extra Copies of ULEs vector is a Gateway compatible *Agrobacterium* sp. binary vector. ULE6, ULE7, ULE30 and ULE22 are inserted. The backbone of the vector originates from the destination vector pMCDC123 (Curtis and Grossniklaus 2003) and confers resistance to BASTA.



Supplemental Figure 6. **Map of the RdDM vector used to methylate ULE6, ULE7, ULE30 and ULE22**

The RNA-directed DNA methylation (RdDM) vector is a Gateway compatible *Agrobacterium* sp. binary vector. ULE6, ULE7, ULE30 and ULE22 are inserted in forward and inverse orientation (blue color). ULEs are separated by a spacer fragment consisting of two introns in opposite orientations, catalase-1 intron of castor bean (cat intron) and pdk intron. ULEs are transcribed under the constitutive promoter CaMV35S. The backbone of the vector originates from the destination vector pHellsgate12.

Supplementary Tables

Supplemental Table 1. BAC clones used as probes for FISH to analyze the pairing frequency of ULE-BAC regions and non-ULE-BAC regions.

Chromosome region	Chromosome	Dig-labeled BAC	Biotin-labeled BAC
ULE6	1	F6F9	F14P1
ULE30	1	F27G20	F5D14
R1_1	1	T2P11	T24P13
R1_2	1	F15H21	F1N19
ULE3	2	F4P9	T1B8
ULE12	2	T16B24	T7F6
R2	2	T6B20	T9D9
ULE4	3	T27C4	T6K12
R3	3	MIGL6	K20I9
ULE25	4	T10C21	F6I18
R4	4	F13C5	T18B16
ULE16	5	T28N17	F20L16
ULE11	5	K19P17	MIJP23
ULE1	5	MTG10	MMI9
R5	5	K6A12	MIPF21

Supplemental Table 2. ULE sequences together with their flanking sites that were inserted in wild type *A. thaliana* plants.

ULE	Sequence
ULE6	TAATTCGTTTAAAGAGTCTAAAGCTGCAACGGCATCGCCATATACAGAAAGTTTTAAGCGC AGGATAAGAGCATGCACACTCTTTCCCATTTCCAAGCAA
ULE7	GAGGTTTTTGGTTCTAGTGGTGAAAGGGATTGTTGGGTACAATGATGGATGTTTCCTACTG AGGAGAAAAGATGATTGGTTATTTTGTCTGAATT
ULE30	AGAAAATCAATGACAAGGTAGTATGTAGTGAATGGTTGTTTCTTTTGTGTGAAGTATATGT GAGAAAATGACACTTGAGTGTGTGTGAGAGAGAGGA
ULE7	TATGCTTCAACGGTGGCGATAGACAATATATATGCCAACCTTTATTACAACTATCACACAA AAGCATCACTTCACCGTGTTTGTCACTTTTACATTGCTCACACGCTTGATCAACCCCTTCTT

Supplemental Table 3. Primers used for the McrBC-PCR assay and size of genomic regions amplified.

McrBC-PCR primers	Sequence	Amplified region
TA2	Forward: CAAGCCTAGTGAAGCTACAAGC	500 bp
	Reverse: CTGCCCAGAACTCTTCTTCT	
ULE6	Forward: GACCAAATTCGACCCTTCAAT	691 bp
	Reverse: AAGAAGCCTCCTCCTTTGTCA	
ULE30	Forward: GCCATGTCGATGATGGTTTAC	662 bp
	Reverse: TGGTTTGAAAACACAAATAAAGGA	
ULE7	Forward: ACGGATCCATTTTTCGAGTGT	697 bp
	Reverse: TCAATTCCTCCCTAGACCAAAA	
ULE22	Forward: GTGAAGCGGTTTGGAGGTTAT	670 bp
	Reverse: GAGGAAAACCACTCCCGTAAA	

CHAPTER 4

General Discussion

In this study ULEs, highly conserved non-coding sequences were identified for the first time in plants. We identified one set of ULEs shared between dicotyledonous plants and another one between monocot plants (grasses). We further showed that ULEs have unique properties and they are under purifying selection indicating that they are indeed functional elements. Surprisingly, the properties of the ULEs are similar to those of the animal non-coding ultraconserved elements (ncUCEs) prompting that both sets of sequences (plant ULEs and animal ncUCEs) are products of convergent evolution and perhaps serve the same function.

During evolution ULEs were depleted from duplications, suggesting that they might be dosage sensitive. Because of the genome-wide distribution and the uniqueness of these elements, we investigated whether ULEs participate in a chromosome copy counting mechanism. We provide evidence that ULE containing regions might be involved in a pairing mechanism since they show elevated chromosome pairing in somatic cells.

Identification of ULEs and pitfalls

Identification of ULEs is largely dependent on the selection of genomes to be compared as well as the parameters used to define a ULE. In our study, we found 36 ULEs between *Arabidopsis* and grapevine which are ~115 Mya apart. In contrast, when the same analysis was performed between closely related species, such as the grasses, *Brachypodium*, rice, maize and sorghum, 870 ULEs were identified. Moreover, these grass species diverged only ~50 Mya.

In our study, ULEs had a conservation and length cutoff of 85% and 55 bp respectively. These parameters, although they seem arbitrary, are strongly reliant to the biological question to be addressed. In our project, we tried to exclude from our findings potential transcription factor binding sites (TFBSs), therefore we queried for longer conserved sequences (>55bp). Longest TFBS is 50 bp. Since we were interested for UCE-like elements our identity threshold was set high enough to 85%. An example of how the number of CNSs is affected by the genome choice and cutoff criteria is the following: selecting genomes of the Brassicaceae family, which are more closely related, 14-20 Mya resulted the identification of 90’000 CNSs of median length 36 bp (Haudry et al. 2013). From these only 3.4% are present in the phylogenetically neighbor genome of papaya (70 Mya of divergence) and just 0.8% are found in rice (125-235 Mya). The majority of them

were small noncoding RNAs. This is in agreement with our results that longer evolutionary distances, conserved sequences with no genic functions are more scarce.

Although genome comparison studies between different species seem a straight forward approach, there are a few things one should take into consideration when genome-wide analysis for CNSs identification is performed. First, the genomes to be compared should be completely or nearly completely sequenced. Despite the fact that the sequence quality of *Arabidopsis* and grapevine is relatively good, this is not the case for the poplar, papaya and cucumber genomes. When we queried whether ULEs are also present in these recently sequenced genomes, it was not true for all of them. It should also be noted that these plants diverged more recently from *Arabidopsis* than grapevine. Further analysis, revealed that the genes neighboring the missing ULEs are also absent from the genomes of poplar and papaya and in the case of cucumber only the intergenic ULEs were present. This suggests that sequencing status of these genomes is far from complete and using them directly in pairwise genomes alignments would only yield to identification of some but not all ULEs.

According to our definition, a ULE should not be part of exons, mitochondrial DNA, or in case of plant genomes not chloroplast DNA and should be devoid from repetitive DNA. In addition, they should not be tRNAs or functional ncRNA (nucleolar RNAs, miRNAs). Thus, in order to make sure that a sequence is a true non-coding one and not a missed exon, it is necessary to compare species whose genomes are accurately annotated. The *Arabidopsis* genome annotation is maybe the best in plants, but during the course of our analysis we came across multiple missed annotated exons and ncRNAs. To minimize this problem, always the latest available genome annotation should be used. In addition, blastx searches against the nonredundant NCBI protein database can be used to identify and eliminate protein sequences that are not annotated.

A small fraction of the genomes is contaminated by bacterial insertions, which originate from the genome sequencing production. Bacterial contamination is exclusively coming from *Escherichia coli* sequences. Therefore, conserved sequences between genomes should be used in blast searches against an *E. coli* database to cull of unwanted artifacts. In a recent study the identification of UCE elements (100% identity) was reported that are present in plants (Reneker et al. 2012). However, our analysis revealed that these sequences are not part of plant genomic sequences but rather the result of *E. coli* contamination.

In animals, a CNS should be located in collinear positions in different species. However, in plants this might always not be the case. In our study, five *Arabidopsis* ULEs were not found in collinear regions relative to the grapevine and even when we examined a 50 Kb region still only the ULE was conserved. Collinearity in gene order erodes much faster in plants than in animals (Wicker et al. 2010). Interestingly, in three instances ULEs were flanked by repeats. We cannot prove whether the TEs caused the movement of the ULEs. However, our findings suggest that when querying for ULEs in plants should take into account that a certain degree of genome reshuffling due to TE activity might cause translocation of a number of them to non-collinear positions.

Interestingly, alignments of 20 angiosperm species including *Arabidopsis*, grapevine and rice did not yield any uninterrupted sequences, like the 100% identity observed in the animal UCEs. This is true even when the cutoff is just 18 bp (Hupalo and Kern 2013). The lifestyle of plants should be also taken into consideration. They are sessile organisms exposed to various environmental stresses and need to exhibit a certain degree of plasticity to cope with the external stimuli. Thus, by allowing some sequence variation plants can easily adapt to the constantly changing environments.

ULEs have a dosage dependent nature

NcUCEs have been found to be depleted from SDs and copy number variants in humans. Moreover, they are present in all diploid chromosomes except for chromosome Y, which is a monosome and chromosome 21, trisomies of which are viable. Thus, ncUCE/ULEs are dosage sensitive and may participate in a chromosome copy counting mechanism via sequence comparison (Derti et al. 2006).

In our study we show that ULEs are depleted from the two segmental duplication (SD) events which occurred in the *Arabidopsis* lineage. This finding is impressive because ULEs predate the occurrence of SDs (24-40 Mya) and cover a large portion of the *Arabidopsis* genome more than 70%.

The apparent dosage sensitivity of the ULEs should be tested in more recent SDs. *Brassica* species share a whole genome triplication event which arose 13-17 Mya (Cheng et al. 2013). Monocot ULEs are also found in one copy in the grass genomes but there is no prove whether there was a selection on them to escape SDs. Maize genome underwent a very recent duplication event, just 10 Mya (Van de Peer et al. 2009). The more recent polyploidization in maize together with the recent triplication in *Brassica* offers the

opportunity to check whether ULEs were again absent from the duplicated regions. In case they are it furthers support dosage dependent nature of the ULEs.

The depletion of ULEs from supernumerary B chromosomes would further support the counting hypothesis. B chromosomes are considered as parasitic DNA and they are not essential for the growth and the development of an organism. B chromosomes are monosomes and they do not pair at meiosis with any of the standard chromosomes. Recently, the sequence of the rye B chromosomes has been determined (Martis et al. 2012). It would be interesting to investigate whether monocot ULEs are absent from these chromosomes as our counting model would suggest.

Evidence for the chromosome copy counting hypothesis

It is thought that chromosomes are restricted to the chromosome territories (CTs) and only in rare occasions interchromosomal interactions occur (Cremer and Cremer 2001). However, it seems that the nuclear architecture is more dynamic than we thought. It was shown that CTs interaction in interphase nuclei human cells are common (Branco and Pombo 2006). In *Arabidopsis* the current notion is that in interphase somatic cells homologous chromosome pairing is rather the exception and efficient homologous pairing occurs at the nucleolus organization regions (NORs) and between loci that have long stretches of tandem repeats (Pecinka et al. 2004; Jovtchev et al. 2011). In our study we provide evidence, although replicates are needed, that chromosome regions that contain ULEs are also places where efficient homologous pairing takes place.

This finding is in agreement with the chromosome copy counting model which proposes that pairing should precede so the homologous ULEs have the opportunity to come closer and compare their sequence. Our findings, also suggest that copy counting mechanism is a constant process which takes place throughout the cell cycle. Perhaps, this constant surveillance is needed all the time to ensure that the two homologous chromosomes are always present in a diploid cell.

The copy counting model would be further supported if similar homologous chromosome pairing occurs at the ULE regions of other plants where ULEs are still conserved. Fluorescence *in situ* hybridization is possible also in the interphase nuclei of grapevine and *Brachypodium* (Giannuzzi et al. 2011; Jenkins and Hasterok 2007).

FISH was performed with bacterial artificial clones (BACs) providing a resolution in the order of ~100 kb. Even though this resolution is sufficient to study chromosome pairing,

it would be informative to develop probes spanning a smaller region around the ULEs. This would allow us to detect whether ULEs preferentially pair first, further supporting the copy counting model. Recently, a new tool called Oligopaints was developed for chromosome visualization with FISH (Beliveau et al. 2012). Pools of fluorescently labeled probes can be produced by PCR allowing the visualization of shorter regions just a few kilobases. Oligopaints technique is appealing because it allows the development of short probes (53 bp), thus further facilitating the entry of the probe into the nucleus. In addition, probes are strand specific making them ideal for visualizing homologous chromosome pairing.

Physical interactions between CNSs have been shown to exist in human leukemia cells (Robyr et al. 2011). Using a chromosome conformation capture approach (4C) genome-wide interaction map of 10 CNSs was investigated. Surprisingly, CNSs are interacting more often with other CNSs either in *cis* on the same chromosome but more often in *trans* with other chromosomes. This result indicates that CNSs function is likely mediated by their interactions with other CNSs. Thus, in agreement with our FISH results, it is plausible the plant ULEs may also interact in *trans* with their homologues.

Interestingly, addition of four extra copies of ULEs increased the occurrence of chromosomal rearrangements although not statistically significant to the control. Even though statistics argue against the involvement of the pairing of ULEs and chromosomal translocations, it is worth noting that in human lymphocytes chromosomes intermingle with each other and the degree of intermingling is correlated with the frequency of chromosomal translocations (Branco and Pombo 2006). It has been also suggested that chromosome contacts or "chromosome kissing" are prone to chromosomal rearrangements (Cavalli 2007).

Recently it was shown that transcription factors and chromatin remodelers bind to the human ncUCEs and it was hypothesized that ncUCEs are overlapping transcription factor binding sites (Viturawong et al. 2013). INO80 is one of the chromatin remodelers that found to bind to ncUCEs. INO80 is a conserved protein across all eukaryotes and is involved in double strand breaks repair via homologous recombination in somatic cells (Attikum et al. 2004; Fritsch et al. 2004; Seeber et al. 2013b). In order for the damaged DNA molecule to be repaired it requires a template, which can be a sister chromatid or the region from homologous chromosome that has enough sequence similarity. Recently, it was shown that INO80 complex is needed for chromatin mobility in *trans* (Seeber et al. 2013a). Intriguingly, INO80 remodelling complex could also function in the copy

counting hypothesis context by bringing together homologous ncUCEs/ULEs in order the counting to occur.

Alteration of the ULE copy number

ULEs are selected to be absent from duplications indicating that they are dosage sensitive elements. We further tested the dosage sensitivity of the ULEs by perturbation of some of them with insertional mutants or by inserting additional copies. According to the copy counting model perturbation of ULEs will have the same outcome to the addition of extra copies of them. In both cases, there are unpaired ULEs which do not participate in a sequence comparison mechanism and this would trigger deleterious events for the plant.

Perturbation or insertion of ULEs did not cause severe phenotypes, indicating that the effect on phenotype of imbalanced copy number of ULEs is subtle. Therefore, we took advantage of the large number of offspring *Arabidopsis* produce which would allow us to detect even small peculiarities. Indeed, perturbation of one ULE resulted in increased transmission efficiency of the mutant through the male. The phenotype persisted one more generation and then disappeared. We argue that *Arabidopsis* adapts fast enough, and make the altered ULE inactive and does not participate in a sequence comparison process. However, for this we lack evidence. Additional insertional lines on the same or/and other ULEs would likely provided more proof whether our hypothesis is correct. Unfortunately, the only available insertional mutants on ULEs are the ones described in this study.

Genome editing technologies provide opportunities to circumvent the lack of disrupted ULEs. Transcription activator-like effector nucleases (TALENs) are chimeric proteins comprising by sequence specific DNA binding domains fused to a nuclease (Boch et al. 2009; Boch 2011). In the CRISPR/Cas system the DNA target sequence is recognized by a customized RNA molecule (Jinek et al. 2012). Both methods have been used to introduce modifications in animals (Sander et al. 2011; Huang et al. 2011; Tesson et al. 2011; Mali et al. 2013) and recently they have been used to target alterations in plants (Li et al. 2012, 2013; Nekrasov et al. 2013). Now it is feasible, to produce deletions in more ULEs with genome editing.

With genome editing we will be able to explore the transmission efficiency of ULE deletion mutants in an aneuploid trisomic background. We did not see an effect by

perturbation of one of them in this sensitized background. However, due to the scarcity of mutant lines, we cannot draw conclusions whether ULEs are involved in pairing and sequence comparison during meiosis. Interestingly, although aneuploidy causes developmental defects and reduced fertility, its effects are not deleterious for cancer cells. Cancer cells are highly aneuploid (Albertson et al. 2003). It would be interesting to test how the mammalian ncUCEs behave in this cell context. According to the copy counting hypothesis the extra unpaired copies of ncUCEs would trigger deleterious consequences in the cell. However, this is not the case in cancer cells since they exhibit unrestricted growth. Hence, in cancer cells there are ways to cope with the repercussions of unpaired ncUCEs.

References

Albertson DG, Collins C, McCormick F, Gray JW. 2003. Chromosome aberrations in solid tumors. *Nat Genet* **34**: 369–76.

Attikum H Van, Fritsch O, Hohn B, Gasser SM, Ansermet QE, Geneva C-. 2004. Recruitment of the INO80 Complex by H2A Phosphorylation Links ATP-Dependent Chromatin Remodeling with DNA Double-Strand Break Repair. *Cell* **119**: 777–788.

Beliveau BJ, Joyce EF, Apostolopoulos N, Yilmaz F, Fonseka CY, McCole RB, Chang Y, Li JB, Senaratne TN, Williams BR, et al. 2012. Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proc Natl Acad Sci U S A* **109**: 21301–6.

Boch J. 2011. TALEs of genome targeting. *Nat Publ Gr* **29**: 135–136.

Boch J, Scholze H, Schornack S, Landgraf A, Hahn S, Kay S, Lahaye T, Nickstadt A, Bonas U. 2009. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**: 1509–12.

Branco MR, Pombo A. 2006. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol* **4**: e138.

Cavalli G. 2007. Chromosome kissing. *Curr Opin Genet Dev* **17**: 443–50.

Cheng F, Mandáková T, Wu J, Xie Q, Lysak M a, Wang X. 2013. Deciphering the diploid ancestral genome of the Mesohexaploid Brassica rapa. *Plant Cell* **25**: 1541–54.

Cremer T, Cremer C. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**: 292–301.

Derti A, Roth FP, Church GM, Wu C. 2006. Mammalian ultraconserved elements are strongly depleted among segmental duplications and copy number variants. *Nat Genet* **38**: 1216–20.

Fritsch O, Benvenuto G, Bowler C, Molinier J, Hohn B. 2004. The INO80 protein controls homologous recombination in Arabidopsis thaliana. *Mol Cell* **16**: 479–85.

Giannuzzi G, D'Addabbo P, Gasparro M, Martinelli M, Carelli FN, Antonacci D, Ventura M. 2011. Analysis of high-identity segmental duplications in the grapevine genome. *BMC Genomics* **12**: 436.

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet* **45**: 891–8.

Huang P, Xiao A, Zhou M, Zhu Z, Lin S, Zhang B. 2011. Heritable gene targeting in zebrafish using customized TALENs. *Nat Biotechnol* **29**: 699–700.

Hupalo D, Kern AD. 2013. Conservation and Functional Element Discovery in 20 Angiosperm Plant Genomes. *Mol Biol Evol* **30**: 1729-44.

Jenkins G, Hasterok R. 2007. BAC "landing" on chromosomes of Brachypodium distachyon for comparative genome alignment. *Nat Protoc* **2**: 88–98.

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna J a, Charpentier E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**: 816–21.

Jovtchev G, Borisova BE, Kuhlmann M, Fuchs J, Watanabe K, Schubert I, Mette MF. 2011. Pairing of lacO tandem repeats in Arabidopsis thaliana nuclei requires the presence of hypermethylated, large arrays at two chromosomal positions, but does not depend on H3-lysine-9-dimethylation. *Chromosoma* **120**: 609–19.

Li J-F, Norville JE, Aach J, McCormack M, Zhang D, Bush J, Church GM, Sheen J. 2013. Multiplex and homologous recombination-mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9. *Nat Biotechnol* **31**: 688–91.

Li T, Liu B, Spalding MH, Weeks DP, Yang B. 2012. High-efficiency TALEN-based gene editing produces disease-resistant rice. *Nat Biotechnol* **30**: 390–2.

Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-guided human genome engineering via Cas9. *Science* **339**: 823–6.

Martis MM, Klemme S, Banaei-Moghaddam AM, Blattner FR, Macas J, Schmutzer T, Scholz U, Gundlach H, Wicker T, Šimková H, et al. 2012. Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proc Natl Acad Sci U S A* **109**: 13343–6.

Nekrasov V, Staskawicz B, Weigel D, Jones JD, Kamoun S. 2013. Targeted mutagenesis in the model plant Nicotiana benthamiana using Cas9 RNA-guided endonuclease. *Nat Biotechnol* **31**: 691–93.

Pecinka A, Schubert V, Meister A, Kreth G, Klatte M, Lysak M a, Fuchs J, Schubert I. 2004. Chromosome territory arrangement and homologous pairing in nuclei of Arabidopsis thaliana are predominantly random except for NOR-bearing chromosomes. *Chromosoma* **113**: 258–69.

Van de Peer Y, Fawcett J a, Proost S, Sterck L, Vandepoele K. 2009. The flowering world: a tale of duplications. *Trends Plant Sci* **14**: 680–8.

Reneker J, Lyons E, Conant GC, Pires JC, Freeling M, Shyu C-R, Korkin D. 2012. Long identical multispecies elements in plant and animal genomes. *Proc Natl Acad Sci* **109**.

Robyr D, Friedli M, Gehrig C, Arcangeli M, Marin M, Guipponi M, Farinelli L, Barde I, Verp S, Trono D, et al. 2011. Chromosome conformation capture uncovers potential genome-wide interactions between human conserved non-coding sequences. *PLoS One* **6**: e17634.

Sander JD, Cade L, Khayter C, Reyon D, Peterson RT, Joung JK, Yeh J-RJ. 2011. Targeted gene disruption in somatic zebrafish cells using engineered TALENs. *Nat Biotechnol* **29**: 697–8.

Seeber A, Dion V, Gasser SM. 2013a. Checkpoint kinases and the INO80 nucleosome remodeling complex enhance global chromatin mobility in response to DNA damage. *Genes Dev* **27**: 1999–2008.

Seeber A, Hauer M, Gasser SM. 2013b. Nucleosome remodelers in double-strand break repair. *Curr Opin Genet Dev* **23**: 174–84.

Tesson L, Usal C, Ménoret S, Leung E, Niles BJ, Remy S, Santiago Y, Vincent AI, Meng X, Zhang L, et al. 2011. Knockout rats generated by embryo microinjection of TALENs. *Nat Biotechnol* **29**: 695–6.

Viturawong T, Meissner F, Butter F, Mann M. 2013. A DNA-Centric Protein Interaction Map of Ultraconserved Elements Reveals Contribution of Transcription Factor Binding Hubs to Conservation. *Cell Rep* 1–15.

Wicker T, Buchmann JP, Keller B. 2010. Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res* **20**: 1229–37.

Appendix

Contribution to other projects

During my PhD thesis I have contributed to the following projects:

- Boisson-Dernier A, Roy S, **Kritsas K,** Grobei MA, Jaciubek M, Schroeder JI, Grossniklaus U. 2009. Disruption of the pollen-expressed *FERONIA* homologs *ANXUR1* and *ANXUR2* triggers pollen tube discharge. *Development* **136**: 3279-88.

- Schauer MA, Schauer SE, **Kritsas K,** Brunner A, Roschitzki B, Wicker T, Grossniklaus U. *Arabidopsis* male gamete proteome: new proteins, genes and patterns for sperm cell biology. Unpublished

- Schauer MA, Qeli E, **Kritsas K,** Roschitzki B, Rehrauer H, Panse C, Ahrens CH, Grossniklaus U. Proteomics of pollen tube development in *Arabidopsis thaliana*. Unpublished

Disruption of the pollen-expressed *FERONIA* homologs *ANXUR1* and *ANXUR2* triggers pollen tube discharge

Boisson-Dernier A, Roy S, **Kritsas K,** Grobei MA, Jaciubek M, Schroeder JI, Grossniklaus U.

Abstract

The precise delivery of male to female gametes during reproduction in eukaryotes requires complex signal exchanges and a flawless communication between male and female tissues. In angiosperms, molecular mechanisms have recently been revealed that are crucial for the dialog between male (pollen tube) and female gametophytes required for successful sperm delivery. When pollen tubes reach the female gametophyte, they arrest growth, burst and discharge their sperm cells. These processes are under the control of the female gametophyte via the receptor-like serine-threonine kinase (RLK) FERONIA (FER). However, the male signaling components that control the sperm delivery remain elusive. Here, we show that ANXUR1 and ANXUR2 (ANX1, ANX2), which encode the closest homologs of the FER-RLK in Arabidopsis, are preferentially expressed in pollen. Moreover, ANX1-YFP and ANX2- YFP fusion proteins display polar localization to the plasma membrane at the tip of the pollen tube. Finally, genetic analyses demonstrate that ANX1 and ANX2 function redundantly to control the timing of pollen tube discharge as anx1 anx2 double- mutant pollen tubes cease their growth and burst in vitro and fail to reach the female gametophytes in vivo. We propose that ANX-RLKs constitutively inhibit pollen tube rupture and sperm discharge at the tip of growing pollen tubes to sustain their growth within maternal tissues until they reach the female gametophytes. Upon arrival, the female FER-dependent signaling cascade is activated to mediate pollen tube reception and fertilization, while male ANX-dependent signaling is deactivated, enabling the pollen tube to rupture and deliver its sperm cells to effect fertilization.

In this work, I did a phylogenetic tree showing the evolutionary relationship of ANX1 and ANX2 homologs in *Arabidopsis thaliana* (mouse-ear cress), *Oryza sativa* (rice), *Cardamine flexuosa* (wavy bittercress), *Brassica oleracea* (wild cabbage), *Populus trichocarpa* (poplar), *Vitis vinifera* (grapevine) and *Ricimus communis* (castor oil plant). Multiple alignments were performed with CLUSTALW. Phylogenetic analysis was

performed with the PHYLIP package using the protein sequence parsimony method (PROTRAPS) on 100 bootstrap replicates with jumbling of the order of sequences three times for each replicate.

Arabidopsis male gamete proteome: new proteins, genes and patterns for sperm cell biology

Schauer MA, Schauer SE, **Kritsas K,** Brunner A, Roschitzki B, Wicker T, Grossniklaus U.

Abstract

The male gametes have evolved to ensure the safe delivery of the paternal genome to the egg. Before reaching maturity, the sperm cells undergo radical biological changes: condensation of its chromatin, the addition of epigenetic marks, and a reduction in cytoplasm. Although less is known about them, plants, like animals, also have sperm cells, which have been proven difficult to isolate due to their inaccessibility inside the pollen grain. Using a novel isolation approach, the large-scale proteome of sperm cells of the model plant *Arabidopsis thaliana*, containing over 1100 proteins, was characterized. Over half of these proteins did not have a corresponding transcript detected in the sperm cell transcriptome. Analysis of the posttranslational modifications led to the identification of patterns of methylation on ribosomal proteins and elongation factors in plant sperm cells. New unannotated genes, encoding protein domains known to play a role in the reproductive dialog in plants, were uncovered using proteogenomics. Comparisons of sperm proteomes of several model organisms revealed protein orthologs associated with the male development across plant and animal kingdoms. Together, the analysis of the *Arabidopsis* sperm cell proteome, posttranslational modifications of sperm identified proteins, proteogenomics, and sperm cell proteome comparisons across model organisms provides new insights into sperm cell biology.

In this work, I used proteogenomics to identify unannotated genes in the *Arabidopsis* genome. The *Arabidopsis* genome was translated in all six-frames using the stop codons: TAG, TAA, TGA and considering open reading frames (ORFs) which are larger than 24 bp. Approximately, 8.5 million ORFS were generated and translated into protein sequences. These ORFs were used as a reference to identify annotated proteins. From this survey new protein models were uncovered, such as proteins that have the plant self-incompatibility protein S1 domain.

Proteomics of pollen tube development in *Arabidopsis thaliana*

Schauer MA, Qeli E, **Kritsas K,** Roschitzki B, Rehrauer H, Panse C, Ahrens CH, Grossniklaus U.

Abstract

In plants, the pollen tube (PT) transports the male gametes to the ovule for fertilization. After pollen germination, the vegetative cell of the pollen grain, which contains the sperm cells, forms a tube-like structure that grows through the female tissues and ensures the delivery of the immobile male gametes to their female partner. Using a proteomic approach, we report a comprehensive protein map containing over 2800 proteins from *Arabidopsis thaliana* PTs. Functional analysis of the proteome highlights the contribution of mitochondrial-related pathways and transporters during PT development. We analyzed the similarities between PT proteome with related *Arabidopsis* large-scale datasets. Our dataset extends the *Arabidopsis* PT proteome by over an order of magnitude, representing the largest proteome dataset reported so far in PT biology. Additionally, proteins were identified that were previously unrelated to male development. The functional analysis of PT proteome allowed us to draw connections between different signaling pathways.

In this work, I compared the protein sequences from mature pollen (MT) and pollen tube (PT) proteomes from different plant species to the *Arabidopsis* MT and PT proteome. I used protein sequences from the angiosperm *Pinus strobes* (pine), the monocots *Oryza sativa* (rice) and *Lilium longiflorum* (Easter lily) and the *Brassica* species *Brassica napa* (canola).

Supplementary Tables from chapter 2

Supplemental Table S1. ULE coordinates on *Arabidopsis thaliana* genome (TAIR9) and sequence identity to *Vitis vinifera*

ULEs	Chr	Coordinates TAIR9	Type of DNA	Identity to V.vinifera	Sequence
ULE1	5	24909657-24909712	Intergenic	89.00%	CCTGCCTGTTCACAGCACGACAAAGCCACTTCCCAATAAAAACACAACACCTTTCC
ULE2	4	5372337-5372403	Intergenic	86.00%	GAATTGGTGCCTTTGGAAGGAGCAGAAGAAGGAACAGGAAGCTTCTTTGGTAGCATTTGTTCCCAGT
ULE3	2	14170126-14170230	Intergenic	91.00%	TGTTGGAATCTCTTAAATAGGGTGTTTATGTTGGGTATGTCATAATTCACATTCAGGAGGGTGATGAC GCTGTCTGCTAAATAAGAAGACTACCAGGTGTAGTGA
ULE4	3	1208156-1208248	Intronic	85.00%	TATCTTCTTCTTAAATTTGTTGTGGATTGGGTGGATGGGCAATTATCTTACCTAGAGGCTCATGAAACA GCCGTTGTCATTAATTTCTGCATG
ULE5	5	23131441-23131499	Intergenic	89.00%	AGTGATGGGTTTTAGACTAGCAAAGAGAATCCATGTTGGGTGCTTCAAAAGGCGAGCCA
ULE6	1	6886635-6886703	Intergenic	89.00%	TAAAGCTGCAACGGCATCGCCATATACAGAAAGTTTTAAGCGCAGGATAAGAGCATGCACACTCTTTCC
ULE7	1	7760756-7760815	Intergenic	86.00%	TGAAAGGGATTGTTGGGTACAATGATGGATGTTTCCTACTGAGGAGAAAAGATGATTGGT
ULE8	3	5247201-5247262	Intergenic	87.00%	ATCAGCAAAAGGAATCATTATCTTTAAACCAAGAAATTAAAGACGCCTTTAATTACATTCCT
ULE9	4	13349363-13349419	Intergenic	85.00%	CTCCAGACCCATCAGCTTTGCAACAACCCCCGTCTTGCACCATACCACACTCTCTCC
ULE10	4	8561109-8561173	Intronic	86.00%	AGGTGAAAGATCTCCAGAAGGCGACAAGAACATTACAGACTCTATGTTCTGAAGCCAAGGTTAGT
ULE11	5	21904336-21904404	Intergenic	87.00%	TGAGTATGCGTGCGTGCATGAGAAAAAGTGCAGTGGTATTCTCTGTCAGTGATGAAAATAAAAGAGAGG
ULE12	2	16339188-16339254	Intergenic	92.00%	GAGGAATCCATGTGTATCCAAGCACTTGCAGACAAAATGAGCTTCAGAGAAGAGAAAAGCAACCTTT
ULE13	3	8244890-8244946	Intergenic	87.00%	CCCAAAACCTGAGGGTTTGCATGTTCCCGACAATGTCTGGGTTTTGAATGGCATGAT
ULE14	3	2578389-2578454	Intergenic	86.00%	ATGATGATAGACTTGTGAGTGAAGAAAAATGTATCGAAAAAGACTGGTTTAGAGTCAAAGACAAGA
ULE15	2	12025411-12025467	Intergenic	94.00%	TTGAGTAAGCAAGTGGTTAGTGTTCATGGTTGCCTTTCATTCACGAAAAGTAAGTGA
ULE16	2	11073052-11073107	Intronic	91.00%	CCTACATACCGTACTGATTGGTGAAATTATTGGGTTACTCCTGCAGAAGCCCATCC
ULE17	5	10010138-	Intronic	87.00%	TTGGGAGCATCAGTTTTTCGGTGGACAGAATGATATCTATACTCGTAAATAATCTTC

		10010203			ACAGTATGT
ULE18	5	6177762-6177820	Intronic	88.00%	ATGGAAGTAAAGAAGGGGAGCATCAAAGAGTTGTGATGTTATCTGATACAATCTTCCTA
ULE19	5	4226312-4226369	Intronic	93.00%	GGATTATATAGCGGGTTGCCATATCGCCCAAGAAGCTGACAATGACATTACCCATAAT
ULE20	2	209953-210009	Intronic	87.00%	AACCAGGTGGGAGGTGTTCATTCACATCTTATCTCTAGAGTCCAGAGGGATAAAAGC
ULE21	2	6358093-6358152	Intronic	91.00%	AAGATGCTTCTTGTCTACCGTTGTGGTGGTTATCTTCTACTGATACCAAGATGCTGTTTT
ULE22	1	23552178-23552261	Intergenic	85.00%	TAGACAATATATATGCCAACCTTTATTACAACTATCACACAAAAGCATCACTTCACCGTGTTTGTCA CTTTTACATTGCTCACA
ULE23	2	10924895-10924962	Intronic	86.00%	TAGAGAAGCACCTCACCAGTAGCTCTTCCCTTAGACCACTCGATTGGAATCTCTAGAATTCTTTCAAA
ULE24	5	23295760-23295820	Intergenic	88.00%	ACCGACAAAGGGTTCAAAGGGTGCCATTAATGGAGGCGACAACACTTGGATTCCCACCTGAT
ULE25	4	14959610-14959676	Intronic	86.00%	ATATTAATCTTTGTCCATATTTACAGCAAGGTCACATTGCTTAAGCGCCCCACCAAGAATTAGGACA
ULE26	5	23934796-23934855	Intergenic	85.00%	TGACATGCGTGTTGTTTTTTGCGCTGGAGACTCTGAAAACATCGCGTGGAGCGGCTGAGA
ULE27 #	2	8997929-8997994	Intergenic	85.00%	TTACAAAGATGATGAGCCGCACCAAGAGCAGCCCAGCTAGAACCCACGACAACAACCTTCTTCTTCTT
ULE28 #	2	8997606-8997671	Intergenic	85.00%	TTACAAAGATGATGAGCCGCACCAAGAGCAGCCCAGCTAGAACCCACGACAACAACCTTCTTCTTC
ULE29	5	12425783-12425839	Intergenic	87.00%	AAGGTAAAGAGTTCATAATGAATTGGACCAAGAAATTCTCATGAACTTCCATTCCCA
ULE30	1	11630637-11630693	Intergenic	98.00%	GTATGTAGTGAATGGTTGTTTCTTTTGTGTGAAGTATATGTGAGAAAATGACACTTG
ULE31	4	15651488-15651554	Intronic	88.00%	AGTACAACCTTAAAACTATATCACATGAAAAGCAGGTGTTGCGATCAGACCTGTATCCTACAGCATA
ULE32	2	16678961-16679053	Intronic	85.00%	AACTACCTTTACTTCATCATCTACCAGAGATCATAACAGATTACGCCCACATTACCTACAATAAAAGATACAT TTTATTTGGAAGCATGAACA
ULE33	1	5229152-5229233	Intronic	86.00%	TATGCACGTGAATGGAAACTTATCTTTGCCTCTATAAAATTGTCAATATGCACACTTTTTGGACCGAAATAAAA AATAACCT
ULE34	5	2597827-2597897	Intronic	85.00%	CGGCGAAGTCTTCTCAATATAAGGGCCATGTGGCAGCCAAATGGTCCTTTGTAAATATGGGCTTCATCAAT
ULE35	5	10964211-10964275	Intergenic	85.00%	TCTATCATATTGCCATGGACTACCCAAAAAGATGACACGCATCCATGGGAATGACATCACACCAA
ULE36	5	1868110-1868174	Intergenic	87.00%	TAGCGCAGCAATGACTCGACACGCTTCCATTAAGCATTTGTGGAAGGCGATCTTAAGGGCTGCGC

: ULE27 and ULE28 are the two paralogous elements

Supplemental Table S2. ULEs conserved in other plant genomes except for *Arabidopsis* and *Vitis*.

ULE	Poplar	Papaya	Cucumber	Rice	Sorghum	Brachy-podium	Maize
ULE1	No	**Yes**	No	No	No	No	No
ULE2	**Yes**	**Yes**	No	No	No	No	No
ULE3	**Yes**	**Yes**	No	**Yes**	**Yes**	**Yes**	**Yes**
ULE4	No	**Yes**	No	No	No	No	No
ULE5	**Yes**	No	No	No	No	No	No
ULE6	No	No	No	No	No	No	No
ULE7	**Yes**	No	**Yes**	No	No	No	No
ULE8	No	No	No	No	No	No	No
ULE9	No	No	No	No	No	No	No
ULE10	**Yes**	No	No	No	No	No	No
ULE11	**Yes**	**Yes**	No	No	No	No	No
ULE12	**Yes**	**Yes**	No	No	No	No	No
ULE13	**Yes**	**Yes**	No	No	No	No	No
ULE14	**Yes**	**Yes**	No	No	No	No	No
ULE15	**Yes**	**Yes**	No	No	No	No	No
ULE16	**Yes**	**Yes**	No	No	No	No	No
ULE17	**Yes**	**Yes**	**Yes**	No	No	No	No
ULE18	**Yes**	**Yes**	**Yes**	No	No	No	No
ULE19	**Yes**	**Yes**	No	No	No	**Yes**	No
ULE20	**Yes**	**Yes**	**Yes**	No	No	No	No
ULE21	**Yes**	**Yes**	**Yes**	No	No	No	No
ULE22	**Yes**	No	No	No	No	No	No
ULE23	**Yes**	No	**Yes**	No	No	No	No
ULE24	No	**Yes**	No	No	No	No	No
ULE25	No	No	No	No	No	No	No
ULE26	No	No	No	No	No	No	No
ULE27	No	No	No	No	No	No	No
ULE28	No	No	No	No	No	No	No
ULE29	No	No	No	No	No	No	No
ULE30	**Yes**	**Yes**	**Yes**	No	No	No	No
ULE31	**Yes**	**Yes**	No	No	No	No	No
ULE32	**Yes**	**Yes**	**Yes**	No	No	No	No
ULE33	**Yes**	**Yes**	**Yes**	No	No	No	No
ULE34	**Yes**	No	No	No	No	No	No
ULE35	No	No	No	No	No	No	No
ULE36	No	No	No	No	No	No	No

Supplemental Table S3. Presence of genes neighboring ULEs which are not found in poplar or/and papaya genomes.

ULEs not present in poplar and papaya			
ULE	Gene closest to ULE	Presence of gene in poplar	Presence of gene in papaya
ULE8	At3g15518	No	No
ULE9	At4g26410	No	No
ULE25	At4g30680	No	No
ULE26	At5g59340	No	No
ULE27	At2g20920	Partially	No
ULE28	At2g20920	Partially	No
ULE29	At5g33075	No	No
ULE35	At5g28931	No	No
ULE36	At5g06170	No	No
ULEs not present in poplar			
ULE	Gene closest to ULE	Presence of gene in poplar	
ULE1	At5g62000	Partially	
ULE4	At3g04490	Partially	
ULE24	At5g57920	Partially	
ULEs not present in papaya			
ULE	Gene closest to ULE		Presence of gene in papaya
ULE5	At5g57123		No
ULE7	At1g22030		No
ULE10	At4g14970		No

ULE22	At1g63490		No
ULE23	At2g25660		Partially
ULE34	At5g08110		No

Supplemental Table S4. Collection of Affymetrix ATH1-array data querying a total of 103 different tissue and cell types of *Arabidopsis thaliana.*

Tissue	Differentiation state	DataSetID	Reference
young_leaf	3	ATGE_10	Schmid M et al. 2005
young_leaf	3	ATGE_10	Schmid M et al. 2005
young_leaf	3	ATGE_10	Schmid M et al. 2005
mature_leaf	4	ATGE_12	Schmid M et al. 2005
mature_leaf	4	ATGE_12	Schmid M et al. 2005
mature_leaf	4	ATGE_12	Schmid M et al. 2005
cotyledon	4	ATGE_1	Schmid M et al. 2005
cotyledon	4	ATGE_1	Schmid M et al. 2005
cotyledon	4	ATGE_1	Schmid M et al. 2005
senescent_leaf	4	ATGE_25	Schmid M et al. 2005
senescent_leaf	4	ATGE_25	Schmid M et al. 2005
senescent_leaf	4	ATGE_25	Schmid M et al. 2005
cauline_leaf	4	ATGE_26	Schmid M et al. 2005
cauline_leaf	4	ATGE_26	Schmid M et al. 2005
cauline_leaf	4	ATGE_26	Schmid M et al. 2005
internode_shoot	4	ATGE_27	Schmid M et al. 2005
internode_shoot	4	ATGE_27	Schmid M et al. 2005
internode_shoot	4	ATGE_27	Schmid M et al. 2005
flower_st6	3	ATGE_29	Schmid M et al. 2005
flower_st6	3	ATGE_29	Schmid M et al. 2005
flower_st6	3	ATGE_29	Schmid M et al. 2005
hypocotyl	4	ATGE_2	Schmid M et al. 2005
hypocotyl	4	ATGE_2	Schmid M et al. 2005
hypocotyl	4	ATGE_2	Schmid M et al. 2005
flower_st9	3	ATGE_31	Schmid M et al. 2005
flower_st9	3	ATGE_31	Schmid M et al. 2005
flower_st9	3	ATGE_31	Schmid M et al. 2005
flower_st11	4	ATGE_32	Schmid M et al. 2005
flower_st11	4	ATGE_32	Schmid M et al. 2005
flower_st11	4	ATGE_32	Schmid M et al. 2005
flower_st12	4	ATGE_33	Schmid M et al. 2005
flower_st12	4	ATGE_33	Schmid M et al. 2005
flower_st12	4	ATGE_33	Schmid M et al. 2005
sepal	4	ATGE_34	Schmid M et al. 2005
sepal	4	ATGE_34	Schmid M et al. 2005
sepal	4	ATGE_34	Schmid M et al. 2005
petal	4	ATGE_35	Schmid M et al. 2005
petal	4	ATGE_35	Schmid M et al. 2005
petal	4	ATGE_35	Schmid M et al. 2005
stamen	3	ATGE_36	Schmid M et al. 2005
stamen	3	ATGE_36	Schmid M et al. 2005
stamen	3	ATGE_36	Schmid M et al. 2005
carpel	3	ATGE_37	Schmid M et al. 2005
carpel	3	ATGE_37	Schmid M et al. 2005
carpel	3	ATGE_37	Schmid M et al. 2005
flower_st15	4	ATGE_39	Schmid M et al. 2005
flower_st15	4	ATGE_39	Schmid M et al. 2005
flower_st15	4	ATGE_39	Schmid M et al. 2005
root_d7	3	ATGE_3	Schmid M et al. 2005

root_d7	3	ATGE_3	Schmid M et al. 2005
root_d7	3	ATGE_3	Schmid M et al. 2005
pedicel	4	ATGE_40	Schmid M et al. 2005
pedicel	4	ATGE_40	Schmid M et al. 2005
pedicel	4	ATGE_40	Schmid M et al. 2005
sepal_st15	4	ATGE_41	Schmid M et al. 2005
sepal_st15	4	ATGE_41	Schmid M et al. 2005
sepal_st15	4	ATGE_41	Schmid M et al. 2005
petal_st15	4	ATGE_42	Schmid M et al. 2005
petal_st15	4	ATGE_42	Schmid M et al. 2005
petal_st15	4	ATGE_42	Schmid M et al. 2005
stamen_st15	3	ATGE_43	Schmid M et al. 2005
stamen_st15	3	ATGE_43	Schmid M et al. 2005
stamen_st15	3	ATGE_43	Schmid M et al. 2005
carpel_st15	3	ATGE_45	Schmid M et al. 2005
carpel_st15	3	ATGE_45	Schmid M et al. 2005
carpel_st15	3	ATGE_45	Schmid M et al. 2005
leaf	4	ATGE_5	Schmid M et al. 2005
leaf	4	ATGE_5	Schmid M et al. 2005
leaf	4	ATGE_5	Schmid M et al. 2005
pollen_Schmid	2	ATGE_73	Schmid M et al. 2005
pollen_Schmid	2	ATGE_73	Schmid M et al. 2005
pollen_Schmid	2	ATGE_73	Schmid M et al. 2005
silique_glob_emb	4	ATGE_76	Schmid M et al. 2005
silique_glob_emb	4	ATGE_76	Schmid M et al. 2005
silique_glob_emb	4	ATGE_76	Schmid M et al. 2005
silique_heart_emb	4	ATGE_77	Schmid M et al. 2005
silique_heart_emb	4	ATGE_77	Schmid M et al. 2005
silique_heart_emb	4	ATGE_77	Schmid M et al. 2005
silique_triang_emb	4	ATGE_78	Schmid M et al. 2005
silique_triang_emb	4	ATGE_78	Schmid M et al. 2005
silique_triang_emb	4	ATGE_78	Schmid M et al. 2005
seed_torpedo	4	ATGE_79	Schmid M et al. 2005
seed_torpedo	4	ATGE_79	Schmid M et al. 2005
seed_torpedo	4	ATGE_79	Schmid M et al. 2005
shoot	2	ATGE_7	Schmid M et al. 2005
shoot	2	ATGE_7	Schmid M et al. 2005
shoot	2	ATGE_7	Schmid M et al. 2005
seed_walk_stick	4	ATGE_81	Schmid M et al. 2005
seed_walk_stick	4	ATGE_81	Schmid M et al. 2005
seed_walk_stick	4	ATGE_81	Schmid M et al. 2005
seed_early_curl_cot	4	ATGE_82	Schmid M et al. 2005
seed_early_curl_cot	4	ATGE_82	Schmid M et al. 2005
seed_early_curl_cot	4	ATGE_82	Schmid M et al. 2005
seed_early_green_cot	4	ATGE_83	Schmid M et al. 2005
seed_early_green_cot	4	ATGE_83	Schmid M et al. 2005
seed_early_green_cot	4	ATGE_83	Schmid M et al. 2005
seed_green_cot	4	ATGE_84	Schmid M et al. 2005
seed_green_cot	4	ATGE_84	Schmid M et al. 2005
seed_green_cot	4	ATGE_84	Schmid M et al. 2005
early_rosette	3	ATGE_87	Schmid M et al. 2005
early_rosette	3	ATGE_87	Schmid M et al. 2005
early_rosette	3	ATGE_87	Schmid M et al. 2005
inflor_shoot	2	ATGE_8	Schmid M et al. 2005
inflor_shoot	2	ATGE_8	Schmid M et al. 2005
inflor_shoot	2	ATGE_8	Schmid M et al. 2005
root_d17	4	ATGE_9	Schmid M et al. 2005
root_d17	4	ATGE_9	Schmid M et al. 2005
root_d17	4	ATGE_9	Schmid M et al. 2005
central_cell	1	NA	Wuest, SE et al. 2010
central_cell	1	NA	Wuest, SE et al. 2010
central_cell	1	NA	Wuest, SE et al. 2010
egg_cell	1	NA	Wuest, SE et al. 2010
egg_cell	1	NA	Wuest, SE et al. 2010
egg_cell	1	NA	Wuest, SE et al. 2010
synergid_cell	2	NA	Wuest, SE et al. 2010

synergid_cell	2	NA	Wuest, SE et al. 2010
synergid_cell	2	NA	Wuest, SE et al. 2010
root_endodermis	3	ArexDB	Birnbaum K et al. 2003
root_endodermis	3	ArexDB	Birnbaum K et al. 2003
root_endodermis	3	ArexDB	Birnbaum K et al. 2003
root_stele	3	ArexDB	Birnbaum K et al. 2003
root_stele	3	ArexDB	Birnbaum K et al. 2003
root_stele	3	ArexDB	Birnbaum K et al. 2003
root_xylem	3	ArexDB	Brady SM et al. 2007
root_xylem	3	ArexDB	Brady SM et al. 2007
root_xylem	3	ArexDB	Brady SM et al. 2007
root_columella	3	ArexDB	Brady SM et al. 2007
root_columella	3	ArexDB	Brady SM et al. 2007
root_columella	3	ArexDB	Brady SM et al. 2007
root_cortex	3	ArexDB	Brady SM et al. 2007
root_cortex	3	ArexDB	Brady SM et al. 2007
root_cortex	3	ArexDB	Brady SM et al. 2007
root_epidermis	3	ArexDB	Brady SM et al. 2007
root_epidermis	3	ArexDB	Brady SM et al. 2007
root_epidermis	3	ArexDB	Brady SM et al. 2007
root_ground_tissue	3	ArexDB	Brady SM et al. 2007
root_ground_tissue	3	ArexDB	Brady SM et al. 2007
root_ground_tissue	3	ArexDB	Brady SM et al. 2007
root_protophloem	3	ArexDB	Brady SM et al. 2007
root_protophloem	3	ArexDB	Brady SM et al. 2007
root_protophloem	3	ArexDB	Brady SM et al. 2007
lateral_root_cap	3	ArexDB	Brady SM et al. 2007
lateral_root_cap	3	ArexDB	Brady SM et al. 2007
lateral_root_cap	3	ArexDB	Brady SM et al. 2007
root_pericycle	3	ArexDB	Brady SM et al. 2007
root_pericycle	3	ArexDB	Brady SM et al. 2007
root_pericycle	3	ArexDB	Brady SM et al. 2007
root_companion_cell	4	ArexDB	Brady SM et al. 2007
root_companion_cell	4	ArexDB	Brady SM et al. 2007
root_companion_cell	4	ArexDB	Brady SM et al. 2007
root_atrichoblast	4	ArexDB	Birnbaum K et al. 2003
root_atrichoblast	4	ArexDB	Birnbaum K et al. 2003
root_atrichoblast	4	ArexDB	Birnbaum K et al. 2003
seedling	3	H_Seedling Rep 1	Borges F et al. 2008
seedling	3	H_Seedling Rep 2	Borges F et al. 2008
seedling	3	H_Seedling Rep 3	Borges F et al. 2008
sperm	1	E_Sperm Rep 1	Borges F et al. 2008
sperm	1	E_Sperm Rep 2	Borges F et al. 2008
sperm	1	E_Sperm Rep 3	Borges F et al. 2008
pollen_Borges	2	E_Pollen Rep 1	Borges F et al. 2008
pollen_Borges	2	H_Pollen Rep 2	Borges F et al. 2008
pollen_Borges	2	E_Pollen Rep 3	Borges F et al. 2008
late_ovules	3	NA	Yu HJ et al. 2005
late_ovules	3	NA	Yu HJ et al. 2005
late_ovules	3	NA	Yu HJ et al. 2005
early_ovules	2	NA	Yu HJ et al. 2005
early_ovules	2	NA	Yu HJ et al. 2005
early_ovules	2	NA	Yu HJ et al. 2005
meristem_Clavata3	1	E-GEOD-13596	Yadav RK et al. 2009
meristem_Clavata3	1	E-GEOD-13596	Yadav RK et al. 2009
meristem_Clavata3	1	E-GEOD-13596	Yadav RK et al. 2009
meristem_Fil	1	E-GEOD-13596	Yadav RK et al. 2009
meristem_Fil	1	E-GEOD-13596	Yadav RK et al. 2009
meristem_Fil	1	E-GEOD-13596	Yadav RK et al. 2009
meristem_Wus	1	E-GEOD-13596	Yadav RK et al. 2009
meristem_Wus	1	E-GEOD-13596	Yadav RK et al. 2009
root_quiescent-center	1	ArexDB	Nawy T et al. 2005
root_quiescent-center	1	ArexDB	Nawy T et al. 2005
root_StageI	1	ArexDB	Birnbaum K et al. 2003
root_StageI	1	ArexDB	Birnbaum K et al. 2003
root_StageI	1	ArexDB	Birnbaum K et al. 2003

root_StageI	1	ArexDB	Birnbaum K et al. 2003
embryo_globular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
embryo_globular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
suspensor_globular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
suspensor_globular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-micropylar_globular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-micropylar_globular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-periph_globular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-periph_globular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-chalaz_globular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-chalaz_globular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-chalaz_globular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_chalaz_globular	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_chalaz_globular	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_globular	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_globular	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
embryo_preglobular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
embryo_preglobular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-microp_preglobular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-microp_preglobular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm_preglobular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm_preglobular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-chalaz_preglobular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-chalaz_preglobular	1	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat-chalaz_preglobular	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat-chalaz_preglobular	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_preglobular	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_preglobular	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
whole_seeds_preglobular	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
whole_seeds_preglobular	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
embryo_linearCot	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
embryo_linearCot	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm_linearCot	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm_linearCot	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-chalaz_linearCot	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-chalaz_linearCot	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_chalaz_linearCot	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_chalaz_linearCot	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_linearCot	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_linearCot	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
embryo_heartEmb	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
embryo_heartEmb	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-microp_heartEmb	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-microp_heartEmb	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-periph_heartEmb	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-periph_heartEmb	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-chalaz_heartEmb	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-chalaz_heartEmb	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-chalaz_heartEmb	2	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_chalaz_heartEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_chalaz_heartEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_chalaz_heartEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_heartEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_heartEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
whole_seeds_heartEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
whole_seeds_heartEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
embryo_greenEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
embryo_greenEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-microp_greenEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-microp_greenEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-periph_greenEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-periph_greenEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010

endosperm-chalaz_greenEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-chalaz_greenEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_chalaz_greenEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_chalaz_greenEmb	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_greenEmb	4	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_greenEmb	4	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
whole_seeds_matureEmb	4	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
whole_seeds_matureEmb	4	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
whole_seeds_globular	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
whole_seeds_globular	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
whole_seeds_linearCot	4	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
whole_seeds_linearCot	4	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-microp_linearCot	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-microp_linearCot	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
embryo_bendingCot	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
embryo_bendingCot	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-periph_bendingCot	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-periph_bendingCot	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-chalaz_bendingCot	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
endosperm-chalaz_bendingCot	3	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_chalaz_bendingCot	4	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_chalaz_bendingCot	4	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_bendingCot	4	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
seedcoat_bendingCot	4	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
whole_seeds_bendingCot	4	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010
whole_seeds_bendingCot	4	GSE12404	Kirkbride-arrays, as used in: Le BH et al. 2010

Supplemental Table S5. *A. thaliana* accessions genomes obtained from the 1001 Arabidopsis thaliana genome project (http://www.1001genomes.org/index.html)

***A. thaliana* accessions**					
Agu-1	ICE106	ICE169	ICE50	ICE97	Qui-0
Bak-2	ICE107	ICE173	ICE60	ICE98	Rue3-1
Bak-7	ICE111	ICE181	ICE61	Istisu-1	Sha
C24	ICE112	ICE1	ICE63	Kastel-1	Star-8
Cdm-0	ICE119	ICE212	ICE70	Koch-1	TueSB30-3
Del-10	ICE120	ICE213	ICE71	Lag2.2	Tuescha9
Dog-4	ICE127	ICE216	ICE72	Leo-1	TueV13
Don-0	ICE130	ICE21	ICE73	Ler-1	TueWa1-2
Est-1	ICE134	ICE226	ICE75	Lerik1-3	Vash-1
Eys15-2	ICE138	ICE228	ICE79	Mer-6	Vie-0
Fei-0	ICE150	ICE29	ICE7	Nemrut-1	WalhaesB4
HKT2.4	ICE152	ICE33	ICE91	Nie1-2	Xan-1
ICE102	ICE153	ICE36	ICE92	Ped-0	Yeg-1
ICE104	ICE163	ICE49	ICE93	Pra-6	

Supplemental Table S6. Single nucleotide polymorphisms (SNPs) in ULEs across 83 resequenced *Arabidopsis* accessions.

ULEs	Accession	Position TAIR9	Chromosome	Bp in Col-0	Bp in new accession
ULE2	Nie1-2	5372339	4	A	G
ULE6	Don-0	6886692	1	C	G
ULE9	ICE120	13349377	4	G	A
ULE11	Bak-7	21904359	5	A	G
ULE11	ICE104	21904359	5	A	G
ULE11	ICE106	21904359	5	A	G
ULE11	ICE112	21904359	5	A	G
ULE11	ICE119	21904359	5	A	G
ULE11	ICE120	21904359	5	A	G
ULE11	ICE60	21904359	5	A	G
ULE11	ICE71	21904359	5	A	G
ULE11	ICE92	21904359	5	A	G
ULE11	ICE93	21904359	5	A	G
ULE11	ICE97	21904359	5	A	G
ULE11	Kastel-1	21904359	5	A	G
ULE11	Lerik1-3	21904359	5	A	G
ULE11	Mer-6	21904359	5	A	G
ULE11	TueWa1-2	21904359	5	A	G
ULE11	Vash-1	21904359	5	A	G
ULE11	Xan-1	21904359	5	A	G
ULE12	Est-1	16339224	2	T	C
ULE13	Bak-7	8244931	3	T	G

ULE13	Kastel-1	8244931	3	T	G
ULE16	Yeg-1	11073091	2	C	T
ULE21	ICE71	6358112	2	G	A
ULE21	ICE102	6358138	2	C	T
ULE22	ICE71	23552181	1	A	G
ULE22	Yeg-1	23552181	1	A	G
ULE22	ICE72	23552232	1	A	G
ULE22	Xan-1	23552232	1	A	G
ULE23	Ped-0	10924899	2	C	A
ULE23	Cdm-0	10924899	2	G	A
ULE23	Don-0	10924899	2	G	A
ULE23	ICE104	10924899	2	G	A
ULE23	ICE73	10924899	2	G	A
ULE23	Mer-6	10924899	2	T	A
ULE23	Qui-0	10924934	2	T	C
ULE24	ICE107	23295765	5	C	G
ULE25	Ped-0	14959610	4	A	G
ULE26	ICE127	23934817	5	C	T
ULE26	ICE130	23934817	5	C	T
ULE26	ICE212	23934817	5	C	T
ULE26	ICE213	23934817	5	C	T
ULE26	ICE49	23934817	5	C	T
ULE26	ICE50	23934817	5	C	T
ULE26	ICE61	23934817	5	C	T
ULE26	ICE71	23934817	5	C	T

ULE26	ICE73	23934817	5	C	T
ULE26	ICE79	23934817	5	C	T
ULE26	Ler-1	23934817	5	C	T
ULE26	Ped-0	23934817	5	C	T
ULE27	Bak-2	8997608	2	A	C
ULE27	ICE181	8997608	2	A	C
ULE27	ICE29	8997608	2	A	C
ULE27	Nemrut-1	8997608	2	A	C
ULE27	Pra-6	8997608	2	A	C
ULE27	Bak-7	8997622	2	C	T
ULE27	Agu-1	8997631	2	A	G
ULE27	Bak-2	8997631	2	A	G
ULE27	Bak-7	8997631	2	A	G
ULE27	Cdm-0	8997631	2	A	G
ULE27	Del-10	8997631	2	A	G
ULE27	Dog-4	8997631	2	A	G
ULE27	Don-0	8997631	2	A	G
ULE27	Eys15-2	8997631	2	A	G
ULE27	Fei-0	8997631	2	A	G
ULE27	HKT2.4	8997631	2	A	G
ULE27	ICE102	8997631	2	A	G
ULE27	ICE104	8997631	2	A	G
ULE27	ICE106	8997631	2	A	G
ULE27	ICE107	8997631	2	A	G
ULE27	ICE111	8997631	2	A	G

ULE27	ICE112	8997631	2	A	G
ULE27	ICE119	8997631	2	A	G
ULE27	ICE120	8997631	2	A	G
ULE27	ICE127	8997631	2	A	G
ULE27	ICE130	8997631	2	A	G
ULE27	ICE134	8997631	2	A	G
ULE27	ICE138	8997631	2	A	G
ULE27	ICE150	8997631	2	A	G
ULE27	ICE152	8997631	2	A	G
ULE27	ICE153	8997631	2	A	G
ULE27	ICE163	8997631	2	A	G
ULE27	ICE169	8997631	2	A	G
ULE27	ICE173	8997631	2	A	G
ULE27	ICE181	8997631	2	A	G
ULE27	ICE1	8997631	2	A	G
ULE27	ICE212	8997631	2	A	G
ULE27	ICE213	8997631	2	A	G
ULE27	ICE216	8997631	2	A	G
ULE27	ICE21	8997631	2	A	G
ULE27	ICE226	8997631	2	A	G
ULE27	ICE228	8997631	2	A	G
ULE27	ICE29	8997631	2	A	G
ULE27	ICE33	8997631	2	A	G
ULE27	ICE36	8997631	2	A	G
ULE27	ICE49	8997631	2	A	G

ULE27	ICE50	8997631	2	A	G
ULE27	ICE60	8997631	2	A	G
ULE27	ICE61	8997631	2	A	G
ULE27	ICE63	8997631	2	A	G
ULE27	ICE70	8997631	2	A	G
ULE27	ICE71	8997631	2	A	G
ULE27	ICE72	8997631	2	A	G
ULE27	ICE73	8997631	2	A	G
ULE27	ICE75	8997631	2	A	G
ULE27	ICE79	8997631	2	A	G
ULE27	ICE7	8997631	2	A	G
ULE27	ICE91	8997631	2	A	G
ULE27	ICE92	8997631	2	A	G
ULE27	ICE93	8997631	2	A	G
ULE27	ICE97	8997631	2	A	G
ULE27	ICE98	8997631	2	A	G
ULE27	Istisu-1	8997631	2	A	G
ULE27	Kastel-1	8997631	2	A	G
ULE27	Koch-1	8997631	2	A	G
ULE27	Lag2.2	8997631	2	A	G
ULE27	Leo-1	8997631	2	A	G
ULE27	Lerik1-3	8997631	2	A	G
ULE27	Mer-6	8997631	2	A	G
ULE27	Nemrut-1	8997631	2	A	G
ULE27	Nie1-2	8997631	2	A	G

ULE27	Ped-0	8997631	2	A	G
ULE27	Pra-6	8997631	2	A	G
ULE27	Qui-0	8997631	2	A	G
ULE27	Rue3-1	8997631	2	A	G
ULE27	Sha	8997631	2	A	G
ULE27	Star-8	8997631	2	A	G
ULE27	TueSB30-3	8997631	2	A	G
ULE27	Tuescha9	8997631	2	A	G
ULE27	TueV13	8997631	2	A	G
ULE27	TueWa1-2	8997631	2	A	G
ULE27	Vash-1	8997631	2	A	G
ULE27	Vie-0	8997631	2	A	G
ULE27	WalhaesB4	8997631	2	A	G
ULE27	Xan-1	8997631	2	A	G
ULE27	Yeg-1	8997631	2	A	G
ULE27	ICE1	8997650	2	C	A
ULE27	Agu-1	8997659	2	C	A
ULE27	Bak-2	8997659	2	C	A
ULE27	Bak-7	8997659	2	C	A
ULE27	C24	8997659	2	C	A
ULE27	Cdm-0	8997659	2	C	A
ULE27	Del-10	8997659	2	C	A
ULE27	Dog-4	8997659	2	C	A
ULE27	Don-0	8997659	2	C	A
ULE27	Eys15-2	8997659	2	C	A

ULE27	Fei-0	8997659	2	C	A
ULE27	HKT2.4	8997659	2	C	A
ULE27	ICE102	8997659	2	C	A
ULE27	ICE104	8997659	2	C	A
ULE27	ICE106	8997659	2	C	A
ULE27	ICE107	8997659	2	C	A
ULE27	ICE111	8997659	2	C	A
ULE27	ICE112	8997659	2	C	A
ULE27	ICE119	8997659	2	C	A
ULE27	ICE120	8997659	2	C	A
ULE27	ICE127	8997659	2	C	A
ULE27	ICE130	8997659	2	C	A
ULE27	ICE134	8997659	2	C	A
ULE27	ICE138	8997659	2	C	A
ULE27	ICE150	8997659	2	C	A
ULE27	ICE152	8997659	2	C	A
ULE27	ICE153	8997659	2	C	A
ULE27	ICE163	8997659	2	C	A
ULE27	ICE169	8997659	2	C	A
ULE27	ICE173	8997659	2	C	A
ULE27	ICE181	8997659	2	C	A
ULE27	ICE1	8997659	2	C	A
ULE27	ICE212	8997659	2	C	A
ULE27	ICE213	8997659	2	C	A
ULE27	ICE216	8997659	2	C	A

ULE27	ICE21	8997659	2	C	A
ULE27	ICE226	8997659	2	C	A
ULE27	ICE228	8997659	2	C	A
ULE27	ICE29	8997659	2	C	A
ULE27	ICE33	8997659	2	C	A
ULE27	ICE36	8997659	2	C	A
ULE27	ICE49	8997659	2	C	A
ULE27	ICE50	8997659	2	C	A
ULE27	ICE60	8997659	2	C	A
ULE27	ICE61	8997659	2	C	A
ULE27	ICE63	8997659	2	C	A
ULE27	ICE70	8997659	2	C	A
ULE27	ICE71	8997659	2	C	A
ULE27	ICE72	8997659	2	C	A
ULE27	ICE73	8997659	2	C	A
ULE27	ICE75	8997659	2	C	A
ULE27	ICE79	8997659	2	C	A
ULE27	ICE7	8997659	2	C	A
ULE27	ICE91	8997659	2	C	A
ULE27	ICE92	8997659	2	C	A
ULE27	ICE97	8997659	2	C	A
ULE27	ICE98	8997659	2	C	A
ULE27	Istisu-1	8997659	2	C	A
ULE27	Kastel-1	8997659	2	C	A
ULE27	Koch-1	8997659	2	C	A

ULE27	Lag2.2	8997659	2	C	A
ULE27	Leo-1	8997659	2	C	A
ULE27	Lerik1-3	8997659	2	C	A
ULE27	Mer-6	8997659	2	C	A
ULE27	Nemrut-1	8997659	2	C	A
ULE27	Nie1-2	8997659	2	C	A
ULE27	Ped-0	8997659	2	C	A
ULE27	Pra-6	8997659	2	C	A
ULE27	Qui-0	8997659	2	C	A
ULE27	Rue3-1	8997659	2	C	A
ULE27	Sha	8997659	2	C	A
ULE27	Star-8	8997659	2	C	A
ULE27	TueSB30-3	8997659	2	C	A
ULE27	Tuescha9	8997659	2	C	A
ULE27	TueV13	8997659	2	C	A
ULE27	TueWa1-2	8997659	2	C	A
ULE27	Vash-1	8997659	2	C	A
ULE27	Vie-0	8997659	2	C	A
ULE27	WalhaesB4	8997659	2	C	A
ULE27	Xan-1	8997659	2	C	A
ULE27	Yeg-1	8997659	2	C	A
ULE27	Eys15-2	8997662	2	C	T
ULE27	ICE130	8997669	2	T	A
ULE27	ICE134	8997669	2	T	A
ULE27	ICE138	8997669	2	T	A

ULE27	ICE150	8997669	2	T	A
ULE27	ICE152	8997669	2	T	A
ULE27	ICE153	8997669	2	T	A
ULE27	ICE70	8997669	2	T	A
ULE27	ICE75	8997669	2	T	A
ULE27	Koch-1	8997669	2	T	A
ULE27	Sha	8997669	2	T	A
ULE29	Bak-2	12425789	5	A	G
ULE29	Bak-7	12425789	5	A	G
ULE29	C24	12425789	5	A	G
ULE29	Cdm-0	12425789	5	A	G
ULE29	Del-10	12425789	5	A	G
ULE29	Eys15-2	12425789	5	A	G
ULE29	HKT2.4	12425789	5	A	G
ULE29	ICE102	12425789	5	A	G
ULE29	ICE104	12425789	5	A	G
ULE29	ICE106	12425789	5	A	G
ULE29	ICE107	12425789	5	A	G
ULE29	ICE112	12425789	5	A	G
ULE29	ICE119	12425789	5	A	G
ULE29	ICE163	12425789	5	A	G
ULE29	ICE1	12425789	5	A	G
ULE29	ICE216	12425789	5	A	G
ULE29	ICE29	12425789	5	A	G
ULE29	ICE33	12425789	5	A	G

ULE29	ICE50	12425789	5	A	G
ULE29	ICE91	12425789	5	A	G
ULE29	ICE93	12425789	5	A	G
ULE29	ICE97	12425789	5	A	G
ULE29	ICE98	12425789	5	A	G
ULE29	Istisu-1	12425789	5	A	G
ULE29	Kastel-1	12425789	5	A	G
ULE29	Koch-1	12425789	5	A	G
ULE29	Lerik1-3	12425789	5	A	G
ULE29	Pra-6	12425789	5	A	G
ULE29	Qui-0	12425789	5	A	G
ULE29	Rue3-1	12425789	5	A	G
ULE29	TueSB30-3	12425789	5	A	G
ULE29	Tuescha9	12425789	5	A	G
ULE29	TueV13	12425789	5	A	G
ULE29	Vash-1	12425789	5	A	G
ULE29	WalhaesB4	12425789	5	A	G
ULE29	Xan-1	12425789	5	A	G
ULE29	Bak-2	12425794	5	T	G
ULE29	Bak-7	12425794	5	T	G
ULE29	Xan-1	12425794	5	T	G
ULE29	C24	12425801	5	T	C
ULE29	ICE1	12425831	5	C	T
ULE29	ICE93	12425832	5	C	A
ULE29	ICE97	12425832	5	C	A

ULE29	ICE98	12425832	5	C	A
ULE29	Istisu-1	12425832	5	C	A
ULE29	Lerik1-3	12425832	5	C	A
ULE29	Rue3-1	12425832	5	C	A
ULE29	ICE33	12425837	5	C	A
ULE30	C24	11630665	1	G	T
ULE30	ICE92	11630665	1	G	T
ULE32	Agu-1	16678972	2	C	A
ULE32	Del-10	16678972	2	C	A
ULE32	Don-0	16678972	2	C	A
ULE32	ICE181	16678972	2	C	A
ULE32	ICE216	16678972	2	C	A
ULE32	ICE49	16678972	2	C	A
ULE32	ICE50	16678972	2	C	A
ULE32	ICE61	16678972	2	C	A
ULE32	ICE79	16678972	2	C	A
ULE32	Ped-0	16678972	2	C	A
ULE32	Vie-0	16678972	2	C	A
ULE32	ICE92	16679008	2	C	T
ULE32	Fei-0	16679047	2	A	C
ULE35	ICE169	10964237	5	A	T
ULE35	ICE173	10964237	5	A	T
ULE35	ICE212	10964237	5	A	T
ULE35	ICE213	10964237	5	A	T
ULE35	ICE226	10964237	5	A	T

ULE35	ICE228	10964237	5	A	T
ULE35	ICE79	10964237	5	A	T
ULE35	Agu-1	10964241	5	G	C
ULE35	ICE216	10964241	5	G	C
ULE35	ICE91	10964241	5	G	C
ULE35	ICE93	10964241	5	G	C
ULE35	ICE97	10964241	5	G	C
ULE35	ICE98	10964241	5	G	C
ULE35	Rue3-1	10964241	5	G	C
ULE35	ICE60	10964245	5	A	G
ULE35	Yeg-1	10964245	5	A	G
ULE35	ICE169	10964249	5	G	A
ULE35	ICE173	10964249	5	G	A
ULE35	ICE212	10964249	5	G	A
ULE35	ICE213	10964249	5	G	A
ULE35	ICE226	10964249	5	G	A
ULE35	ICE228	10964249	5	G	A
ULE35	ICE79	10964249	5	G	A
ULE35	Istisu-1	10964249	5	G	A
ULE35	Lerik1-3	10964249	5	G	A
ULE36	Dog-4	1868112	5	G	C
ULE36	Kastel-1	1868112	5	G	C

Supplemental Table S7. Single-base-pair DNA methylation pattern of ULEs in Col-0 ecotype.

ULEs	Methylation	Chr	Position_ TAIR9	Strand	Context	Percent of read that showed methylated cytosine
ULE1	No					
ULE2	No					
ULE3	No					
ULE4	No					
ULE5	No					
ULE6	No					
ULE7	No					
ULE8	No					
ULE9	No					
ULE10	No					
ULE11	No					
ULE12	No					
ULE13	No					
ULE14	No					
ULE15	No					
ULE16	Yes	2	11073060	+	CG	100
ULE16	Yes	2	11073061	-	CG	90
ULE17	Yes	5	10010156	+	CG	90
ULE17	Yes	5	10010157	-	CG	95
ULE17	Yes	5	10010181	+	CG	100
ULE17	Yes	5	10010182	-	CG	100
ULE18	No					

ULE19	No					
ULE20	No					
ULE21	No					
ULE22	Yes	1	23252235	-	CG	79
ULE23	Yes	2	10924911	-	CHG	24
ULE23	Yes	2	10924935	+	CG	83
ULE23	Yes	2	10924936	-	CG	85
ULE24	No					
ULE25	No					
ULE26	No					
ULE27	Yes	2	8997932	+	CHH	33
ULE27	Yes	2	8997944	-	CHH	33
ULE27	Yes	2	8997945	+	CHG	40
ULE27	Yes	2	8997946	+	CG	100
ULE27	Yes	2	8997947	-	CG	56
ULE27	Yes	2	8997950	+	CHH	50
ULE27	Yes	2	8997957	+	CHG	50
ULE27	Yes	2	8997959	-	CHG	43
ULE27	Yes	2	8997964	-	CHG	89
ULE27	Yes	2	8997968	-	CHH	42
ULE27	Yes	2	8997975	+	CG	100
ULE27	Yes	2	8997976	-	CG	73
ULE28	Yes	2	8997609	+	CHH	53
ULE28	Yes	2	8997619	-	CHH	67
ULE28	Yes	2	8997621	-	CHH	50
ULE28	Yes	2	8997622	+	CG	90

ULE28	Yes	2	8997623	-	CG	67
ULE28	Yes	2	8997624	-	CHG	100
ULE28	Yes	2	8997634	+	CHG	40
ULE28	Yes	2	8997635	-	CHG	40
ULE28	Yes	2	8997638	+	CHG	60
ULE28	Yes	2	8997640	-	CHG	100
ULE28	Yes	2	8997644	-	CHH	90
ULE28	Yes	2	8997651	+	CG	100
ULE28	Yes	2	8997652	-	CG	90
ULE29	No					
ULE30	No					
ULE31	No					
ULE32	No					
ULE33	No					
ULE34	Yes	5	2597827	+	CG	100
ULE34	Yes	5	2597828	-	CG	100
ULE34	Yes	5	2597830	+	CG	100
ULE34	Yes	5	2597831	-	CG	91
ULE35	Yes	5	10964250	+	CG	100
ULE35	Yes	5	10964251	-	CG	88
ULE36	Yes	5	1868122	-	CHH	17

Supplemental Table S9. Filters applied on conserved sequences between (A) *Arabidopsis* and *Vitis* and (B) between *Brachypodium* and rice.

A)

Filters	Number of sequences found
Conserved sequences ≥85%, >55 bp	143,861
Coding sequences, proteins TAIR9	48,429
Repeats	73,378
tRNA, rDNA	12,511
Mitochondrial, chloroplast DNA	850
Small non-coding RNAs	4,002
E.coli contamination	0
Sequences belonging to overlapping fragments and are found in more than 6 copies in the genome	4,462
Manual evaluation against *Arabidopsis* Transcriptome Genomic Express Database, BlastX searches	193
ULEs found in *Arabidopsis* - *Vitis*	36

B)

Filters	Number of sequences found
Conserved sequences ≥85%, >55 bp	1,723,486
Coding sequences, proteins	579,427
Repeats	1,093,055
tRNA, rDNA	16,618
Mitochondrial, chloroplast DNA	9,063
Small non-coding RNAs	2,576
E.coli contamination	0
Sequences belonging to overlapping fragments and are found in more than 6 copies in the genome	18,256
Brachypodium – rice conserved sequences	4,572
Monocot ULEs in *Brachypodium*, rice, maize and Sorghum	870

Acknowledgements

It was a real pleasure doing my PhD studies at the Institute of Plant Biology in Zurich. There are so many people who helped me during this journey whom I would really like to thank.

I start with Ueli my supervisor because he gave me the opportunity to work in such an exciting project, for having his door open for discussions, his support, giving me the freedom to learn as much as I wanted and for the beers we had together. Thomas my bioinformatic supervisor, first of all for his warm heart, his optimistic attitude, great humor and for being always there to support and answer my questions. I would like to thank the people from the Bioinformatics lab, Simone, Margarita, James and especially Jan for the amazing team spirit we had and for being such great friends, in fact the best.

I would like to thank Celia for her help with FISH and the fruitful discussions we had during my work at the bench. I thank Aurelien for the constructive input on the project and his positive nature. Valeria for her enduring support in the lab. Arturo and Daniela for the technical help and their friendship. Christof and Afif for microscopy support.

Aurelien, Johan, Mayank, Nuno, Rita, Marek and Sibylle my very good friends in the lab. Daniela and Christian H. for our coffee breaks and their efforts to teach me German. During my PhD I met so many amazing people whom I feel so lucky that I came across: my favorite lab mates: Sharon, Heike, Joana, Roger, Fred, and Valeria; my PhD buddies: Christian S., and Michi; Afif, Hannes, Arturo, Deborah, Daniela, Anja H., Milka, Marc, Lena, Anna, Francis, Lisi, Stefan G., Wanhui, Marian, Arco, Evelyne, Stefan R. and Caroline (cafeteria). Many people to thank but too important people! I will always cherish the time we spent at the institute.

I would like also to thank Ting for being my external supervisor, for always replying to my emails with great enthusiasm, and providing thought provoking feedback.

For the end I kept two very important people in my life, my parents Vaso and Dimitri for their constant support and love.